A NON-ASYMPTOTIC STUDY OF LOW-RANK ESTIMATION OF SMOOTH KERNELS ON GRAPHS

A Thesis
Presented to
The Academic Faculty

by

Pedro A. Rangel

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics

Georgia Institute of Technology
December 2014

A NON-ASYMPTOTIC STUDY OF LOW-RANK ESTIMATION OF SMOOTH KERNELS ON GRAPHS

Approved by:

Professor Vladimir Koltchinskii,
Advisor
School of Mathematics
Georgia Institute of Technology

Professor Anton Leykin
School of Mathematics
Georgia Institute of Technology

Professor Karim Lounici
School of Mathematics
Georgia Institute of Technology

Professor Le Song
College of Computing
Georgia Institute of Technology

Professor Greg Blekherman
School of Mathematics
Georgia Institute of Technology

Date Approved: 21 July 2014

To Erika,

the lighthouse of my life.

You loved me even when I did not deserve it.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of all those who helped me during my time as a graduate student. My deepest gratitude goes to my adviser Vladimir Kotchinskii. Without Vladimir's patience and guidance, I would have been completely hopeless. Not only did he introduce me to interesting questions in mathematics, but also he shed light on all the mathematical difficulties that I faced during my studies. I am also extremely grateful to Professor Karim Lounici for listening to my research ideas and for giving me helpful feedback. And I cannot go without mentioning my academic big brother Stanislav Minsker, who I thank for the many insightful discussions we had.

The School of Mathematics at Georgia Tech fosters a perfect environment to create math, and for that I must thank them profusely. Without their hard work and kindness, graduate student life would have been even harder. I also want to express my gratitude to the National Science Foundation for supporting my research through grants CCF-0808863 and DMS-1207808.

During times of madness, my friends helped me keep my sanity. In particular, without Erika and Alejandro, I would certainly have been in an asylum. Important also are those who were kind enough to listen to my many nonsensical ramblings. Arguably, those ramblings were the core of my graduate student experience. Instead of including an insanely long list of names, I will make sure to thank these patient souls in person.

Last, but not least, I want to thank my parents for their unwavering support. They gave me all the tools I needed, and they always encouraged me to become the person I wanted to be.

Contents

List of Tables

List of Figures

SUMMARY

This dissertation investigates the problem of estimating a kernel over a large graph based on a sample of noisy observations of linear measurements of the kernel. We are interested in solving this estimation problem in the case when the sample size is much smaller than the ambient dimension of the kernel. As is typical in high-dimensional statistics, we are able to design a suitable estimator based on a small number of samples only when the target kernel belongs to a subset of restricted complexity. In our study, we restrict the complexity by considering scenarios where the target kernel is both low-rank and smooth over a graph. The motivations for studying such problems come from various real-world applications like recommender systems and social network analysis.

In the first part, we study the problem of estimating similarity kernels on graphs by employing a modified least squares method with a complexity penalization involving both the nuclear norm and Sobolev-type norm. There are two main contributions in this first part: 1) we introduce a low-coherence function which measures the amount of information that we obtain from a random sample of a kernel on a graph; 2) we prove upper bounds on L_2-type errors of such estimators with explicit dependence on both the rank and the degree of smoothness of the target kernel. The upper bound shows that the proposed estimator requires less samples than standard matrix completion techniques in scenarios where a matrix is naturally indexed by a graph. In particular, the proposed estimator could be used for the problem of predicting links in a social network.

In the second part, we study a more general problem of estimating smooth kernels on graphs. Using standard tools of non-parametric estimation, we derive a minimax

lower bound on the L_2-error in terms of the rank and the degree of smoothness of the target kernel. To prove the optimality of our lower-bound, we proceed to develop upper bounds on the L_2-error for a least-square estimator based on a non-convex penalty. The proof of these upper bounds depends on bounds for estimators over uniformly bounded function classes in terms of Rademacher complexities. We also propose a computationally tractable estimator based on least-squares with convex penalty. We derive an upper bound for the computationally tractable estimator in terms of the coherence function introduced in the first part. Finally, we present some scenarios wherein this upper bound achieves a near-optimal rate.

Chapter I

RECOMMENDER SYSTEMS AND MATRIX COMPLETION

A retail company, hoping to increase its sales, hires us to implement a *recommender system*. That is, a system that accurately predicts the rating that a user would give to an item in an inventory based on previously known ratings [31, 56]. We tackle the problem from a matrix completion perspective where our goal is to predict the blanks of an incomplete utility matrix indexed by users and items. A known entry of this utility matrix contains a value that represents what is known about the degree of preference of that user for that item. Filling the missing values at random completes the matrix without giving any real information about future ratings. Therefore, to achieve meaningful predictions, we assume that few characteristics determine what items a user likes. Based on this heuristic, we are interested in finding a low rank matrix that agrees with our observations.

1.1 Noiseless Low-Rank Matrix Completion

As a first approach, we model the recommender system problem in the *noiseless low-rank matrix completion* setting [10, 13, 14, 32, 54]. In this framework, our goal is to recover an unknown low-rank objective $m_1 \times m_2$ matrix M_* from n observations $M_*(i_1, j_1), \ldots, M_*(i_n, j_n)$ of its entries. We assume that the indexes $(i_1, j_1), \ldots, (i_n, j_n)$ are picked independently and uniformly from the set $\{1, \ldots, m_1\} \times \{1, \ldots, m_2\}$ of indexes of M_*, and that our observations are not corrupted by noise.

For a complex-valued $m_1 \times m_2$ matrix M, let $\text{rank}(M)$ denote the rank of M, M^* denote its adjoint and M^T denote its transpose. In the case when M is a square

matrix, that is when $m_1 = m_2$, we denote its trace by $\mathrm{trace}(M)$. By *singular value decomposition*, there are orthonormal bases $\{u_1, \ldots, u_{m_1}\} \subseteq \mathbb{C}^{m_1}$, $\{u_1, \ldots, u_{m_2}\} \subseteq \mathbb{C}^{m_2}$, and non-negative real numbers $\sigma_1 \geq \cdots \geq \sigma_r$ such that $M = \sum_{k=1}^{r} \sigma_k (u_k \otimes v_k)$, where r is the rank of M. The vectors $u_1, \ldots, u_{m_1}$ and $v_1, \ldots, v_{m_1}$ are called *left and right singular vectors* respectively, while the non-negative real numbers $\sigma_1, \ldots, \sigma_r$ are called *singular values*. Note that we follow the standard convention of ordering the singular values decreasingly.

In the study of matrix completion problems, we mainly use three different matrix norms. The *spectral norm* $\|M\| := \sigma_1$, the *nuclear norm* $\|M\|_* := \sum_{k=1}^{r} \sigma_k$ and the *Frobenius norm* $\|M\|_F^2 := \sum_{k=1}^{r} \sigma_k^2$. We define the *Hilbert-Schmidt* inner product between two $m_1 \times m_2$ matrices M_1 and M_2 as

$$\langle M_1, M_2 \rangle := \mathrm{trace}(M_1 M_2^*)$$

The Frobenius norm turns out to be the norm induced by the Hilbert-Schmidt inner product.

For the sake of simplicity, in this section, we restrict our presentation to the case where the target matrix M_* belongs to the space of hermitian matrices $\mathbb{H}_m$ of size $m \times m$. The most general case follows by hermitian dilation [52, 8]. The *spectral representation* of M has the form $M = \sum_{k=1}^{r} \lambda_k (u_k \otimes u_k)$, where $r = \mathrm{rank}(S)$, $\lambda_1 \leq \cdots \leq \lambda_r$ are non-zero eigenvalues of S repeated with their multiplicities; and $u_1, \ldots, u_r$ are the corresponding orthonormal eigenfunctions. Note that, in the case of repeated eigenvalues, the choice of the eigenfunctions u_js is not unique. Also note that, unlike the singular values, we order the eigenvalues of an Hermitian matrix increasingly. We extend any real function f to the space of hermitian matrices by the usual "functional calculus", that is $f(M) := \sum_{k=1}^{r} f(\lambda_k)(u_k \otimes u_k)$. For hermitian matrices, we can calculate the matrix norms by:

$$\|M\| = \max_{k=1,\ldots,r} |\lambda_k|, \qquad \|M\|_* = \sum_{k=1}^{r} |\lambda_k|, \qquad \|M\|_F^2 = \sum_{k=1}^{r} \lambda_k^2.$$

As a means to exemplify the difficulties of noiseless low-rank matrix completion, let us consider the case where our target matrix M_* has one entry equal to 1 and all the other entries equal to 0. Then $\text{rank}(M_*) = 1$ but the probability that the only nonzero entry is not present in the sample is $\left(1 - \frac{1}{m^2}\right)^n$, which is close to 1 when $n = o(m^2)$. It is therefore impossible to recover an arbitrary low-rank matrices from a "small" set of sampled entries unless we restrict our search to a certain subclass of low-rank matrices. Some of the approaches to restrict the space of matrices include the use of low-coherence assumptions [13, 14, 29], spikeness [47, 25], and genericity [34].

1.1.1 Low-rank matrix completion under low coherence assumptions

In this presentation, we restrict the objective matrix using low-coherence assumptions. The coherence coefficient of an r-rank $m \times m$ matrix is a number ν between 1 and m/r that, roughly speaking, measures how much information a random entry of the matrix can give us. One can check that the coherence constant of an $m \times m$ matrix with one entry equals to 1 and all the other entries equal to 0 is m. In contrast, for instance, the coherence constants of the 256×256 matrices of rank 40 shown in figures 1b and 2b are 2.7 and 2.1 respectively.

As we will see, the number of samples needed to recover M_* depends linearly on its coherence coefficient with respect to the standard basis. To be precise, let U be the range of M and let P_U be the orthogonal projection to U. The *coherence coefficient* for the matrix M with respect to a basis $\{\phi_1, \dots, \phi_m\} \subseteq \mathbb{C}^m$ is the smallest constant ν satisfying:

$$\begin{aligned} \|P_U \phi_i\|_2^2 &\leq \nu \frac{r}{m}, \ i = 1, \dots, m \\ \left|\langle \text{sign}(M)\phi_i, \phi_j \rangle_2\right|^2 &\leq \nu \frac{r}{m^2}, \ i, j = 1, \dots, m \end{aligned} \tag{1}$$

where $\langle \cdot, \cdot \rangle_2$ is the standard inner product in $\mathbb{C}^m$, and $\| \cdot \|_2$ is the euclidean norm.

With enough observations, we might hope that there is only one low-rank matrix

matching the known entries. If this were the case, we would recover the target matrix by solving the optimization problem

$$M_R = \arg\min\{\text{rank}(M) : M(i_k, j_k) = M_*(i_k, j_k)\}$$

This optimization problem is a common sense approach which simply seeks the simplest explanation fitting the observed data. As a matter a fact, with enough observations, this estimator returns the target matrix correctly. Unfortunately, all known algorithms which calculate this estimator precisely require time doubly exponential in the dimension m of the matrix. A tractable approach for matrix completion is based in convex relaxation of the rank minimization problem. The main idea is to substitute rank by its convex envelope over the matrices with bounded spectral norm. This convex envelope turns out to be the nuclear norm $\|M\|_*$ of a matrix M [13]. As a result, we define the estimator M_N as the solution of the following convex optimization problem:

$$M_N = \text{argmin}\{\|M\|_* : M(i_k, j_k) = M_*(i_k, j_k)\}$$

The following highly nontrivial result was proved originally by Candes and Tao using an involved combinatorial argument [14]. The version stated here is an improvement due to Gross [29] with a great simplification of the proof. The ingenuity of Gross argument lies in the use of non-commutative Bernstein inequalities to bound certain stochastic error. The theorem shows that target matrices of "low coherence" can be recovered exactly using the nuclear norm minimization algorithm provided that the number of observed entries is of the order mr up to a log factor.

Theorem 1.1 (Candes and Tao [14], Gross [29])**.** Let ν be the coherence of the target matrix M_* with respect to the standard basis and $C > 0$ a numerical constant. If $n \geq C\nu rm \log^2(m)$, then $M_N = M_*$ with probability at least $1 - m^{-2}$.

1.1.2 Algorithmic considerations

A large amount of research has been devoted to develop improved algorithms for nuclear norm minimization and matrix completion. In [54], Recht, Fazel and Parrilo posted the nuclear norm optimization problem as the following semidefinite programming optimization problem,

$$\begin{aligned} \min_{M,W_1,W_2} \quad & \operatorname{trace}(W_1) + \operatorname{trace}(W_2) \\ \text{subject to:} \quad & \begin{pmatrix} W_1 & M \\ M^T & W_2 \end{pmatrix} \succeq 0, \qquad M(i_k, j_k) = M_*(i_k, j_k), \\ & \qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, n. \end{aligned}$$

Although theoretically interesting, semidefinite programming has a complexity of $O(m^6)$ for $m \times m$-matrices, which is unbearable for large scale applications. This has encouraged efforts to find more efficient algorithms that perform well in practice. In [10], Cai, Candes and Shen propose an iterative singular value thresholding algorithm which does not require the rank to be specified and iteratively optimizes an approximation of the nuclear-norm objective function. In [28], Goldfarb, Ma and Wen analyze a fixed point continuation algorithm for nuclear norm minimization that incorporates an approximate singular value decomposition procedure. In [61], Wen, Yin and Zhang introduce a Low-Rank Matrix Fitting algorithm which fixes the rank by explicitly writing the matrix in terms of its low-rank factors and uses an optimization technique based on successive over-relaxation to minimize the error. In [50], Ngo and Saad present an algorithm that re-interprets Low-rank Matrix Fitting as optimization on the Grassmann manifold and then improves convergence by changing the metric on the manifold and using conjugate gradients rather than standard gradient descent. In [4], Balzano, Nowak, and Recht study a Grassmannian Rank-One Update Subspace Estimation algorithm for tracking subspaces from highly incomplete observations with applications to matrix completion in the case where the observations arrive on-line. In [16], Dai and Milenkovic design an optimization algorithm based

on the observation that matrix completion can be solved by searching for a column space that matches the observations.

1.1.3 An example in image processing

Greyscale digital images are usually encoded as matrices where each entry represents a pixel, and each entry value represents the intensity of the corresponding pixel. In this section, to exemplify the power of matrix completion, we consider the problem of recovering a grayscale digital image from some observations of its pixels. In figures 1 and 2, we present the matrix completion recovery results for two different 256×256 images. Before sampling from the images, we restricted their rank to 40 using singular value thresholding. The figures show both the mask and the recovered images for two different cases, first for a mask containing half of the pixels and second for a mask containing 30% of the pixels. For these examples, we implement the singular value thresholding algorithm introduced by Cai, Candes and Shen in [10].

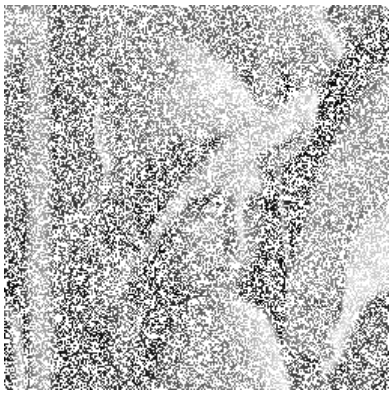
1.2 Noisy Matrix Completion

We proceed to consider the more realistic scenario where the observed entries of the matrix are contaminated with additive zero mean noise [11, 12, 33, 40]. In this case, we do not hope to recover the target matrix exactly, but instead we are interested in designing a statistical estimator that approximates the target matrix accurately with high probability. We measure the performance of an estimator by a norm of its error.

In the analysis of statistical estimators, we are interested in two kinds of results. First, we are interested in finding lower bounds for the best possible error that any estimator can achieve. Second, we are interested in designing estimators and measuring their performance through probabilistic upper bounds on their error. Our final goal is to design computationally tractable estimators with upper bounded error of the same order given by the theoretical limitation imposed by the lower bound. We exemplify this methodology in the analysis of an estimator for a matrix completion

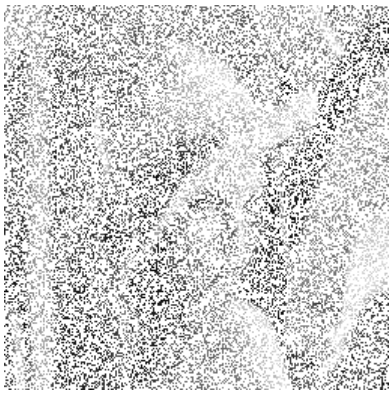
(a) Original Lenna Image

(b) Rank 40 Lenna image



(c) 50% masked Rank 40 image

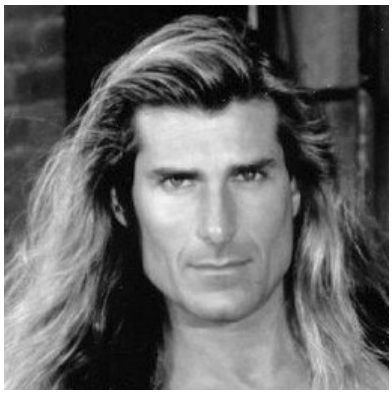
(d) Image recovered from 50% mask



(e) 30% masked Rank 40 image

(f) Image recovered from 30% mask

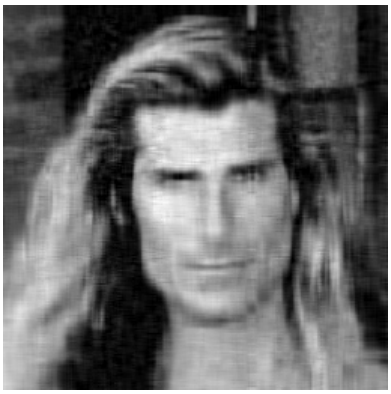
Figure 1: Recovering Lenna through low-rank matrix completion



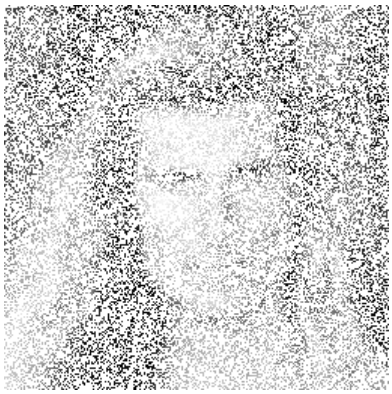
(a) Original Fabio Image

(b) Rank 40 Fabio Image

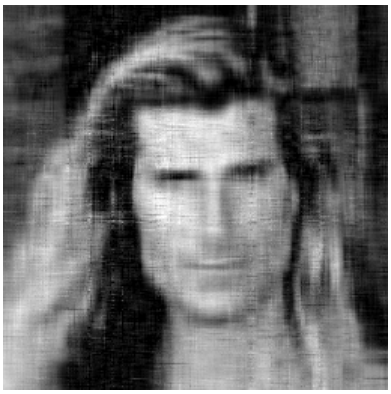
(c) 50% masked Rank 40 image



(d) Image recovered from 50% mask



(e) 30% masked Rank 40 image



(f) Image recovered from 30% mask

Figure 2: Recovering Fabio through low-rank matrix completion

problem with observations contaminated with additive gaussian noise.

1.2.1 Lower bounds for noisy low-rank matrix completion

To be precise, we consider the problem of estimating a matrix M_* in the set $\mathcal{M}(r, a)$ of real-valued r-rank $m \times m$ matrices with entries bounded by a. To solve this task, we have access to noisy observations of its entries $y_k = M_*(i_k, j_k) + \eta_k$, $k = 1, \ldots, n$, where η_k, $k = 1, \ldots, n$ are independent zero-mean Gaussian random variables with variance σ^2.

In this scenario, by standard techniques in the analysis of non-parametric estimators, we can prove the following lower bound on the Frobenious norm error,

Theorem 1.2 (Koltchinskii, Lounici, Tsybakov [40])**.** If $n \geq rm$, then the following bound holds for absolute constants $\beta \in (0, 1)$ and $c > 0$,

$$\inf_{\hat{M}} \sup_{M_* \in \mathcal{M}(r,a)} \mathbb{P}_{M_*} \left(\frac{1}{m^2} \|M_* - \hat{M}\|_F^2 > c(\sigma \wedge a)^2 \frac{mr}{n} \right) \geq \beta$$

where $\inf_{\hat{M}}$ denotes the infimum over all estimators $\hat{M}$ with values in $\mathbb{R}^{m \times m}$, $\mathbb{P}_{M_*}$ denotes the probability of the observations given that the objective matrix is M_*.

Proof. The proof is based on classical lower bounds for non-parametric estimators using Kullback-Leibler divergence. We define the *Kullback-Leibler divergence* of distributions P and Q, with P absolutely continuous with respect to Q (denoted by $P \ll Q$), as

$$K(P\|Q) := \mathbb{E}_P \log \left(\frac{dP}{dQ} \right)$$

where $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivate of P with respect to Q, and $\mathbb{E}_P$ denotes expected value with respect to distribution P. The proof follows from an application of theorem 2.7 in [59]. Here, we present a version of the theorem adapted for the case of matrix completion problems,

Theorem 1.3. If there exists $M_0, \ldots, M_L$ matrices in a subset $\mathcal{M}$ satisfying the following conditions,

1. Each distribution P_{M_i} is absolutely continuous with respect to P_{M_0}.

2. The following inequality holds

$$\frac{1}{L}\sum_{l=1}^{L} K(P_{M_l}\|P_{M_0}) \leq \frac{1}{4}\log(L)$$

3. There is an $s>0$ such that $\|M_i - M_j\|_F^2 \geq s$, for each $i \neq j$, $i,j = 0,\ldots,L$.

then

$$\inf_{\hat{M}} \sup_{M_* \in \mathcal{M}} \mathbb{P}_{M_*}\left(\|M_* - \hat{M}\|_F^2 > c\cdot s\right) \geq \beta$$

where $c>0$ and $\beta>0$ are absolute constants.

We proceed by finding an appropriate collection of matrices satisfying the conditions of the theorem. Let $\tilde{\mathcal{M}}_k$ be the collection of $m\times r$-matrices with entries $+k$ and $-k$, where $k>0$ is a real number that we will pick later. Due to the Varshamov-Gilbert bound (see lemma 2.7 in [59]), there is a subset $\breve{\mathcal{M}}_k \subseteq \tilde{\mathcal{M}}_k$ such that $|\breve{\mathcal{M}}_k| = 2^{rm/8}$ and two different matrices have at least $rm/8$ different elements. Let $\hat{\mathcal{M}}$ be the set of $m\times m$-matrices formed by repeating a matrix $\breve{M}\in\breve{\mathcal{M}}_k$ or by the zero $m\times m$-matrix. To be precise, if $\hat{M}\in\hat{\mathcal{M}}_k$ then $\hat{M}$ is either the zero matrix or it has the form:

$$\hat{M} = \big(\underbrace{\breve{M}|\ldots|\breve{M}}_{\lfloor\frac{m}{r}\rfloor\text{-times}}|O_{m,m-r\lfloor\frac{m}{r}\rfloor}\big)$$

where $\breve{M}$ is a matrix on $\breve{\mathcal{M}}_k$, and O_{m_1,m_2} is the zero $m_1\times m_2$-matrix. By construction each matrix in $\breve{\mathcal{M}}$ has rank r and entries bounded by k, and therefore $\breve{\mathcal{M}}$ is a finite subset of $\mathcal{M}(r,a)$ whenever $k\leq a$.

Let M_0 be the zero $m\times m$-matrix. Taking into account that $\eta_1,\ldots,\eta_n$ are i.i.d zero mean gaussian random variables with variance σ^2, we conclude that for each nonzero matrix $\breve{M}$ in $\breve{\mathcal{M}}$, $P_{M_0}\ll P_{\breve{M}}$ and moreover

$$K(P_{M_0}\|P_{\breve{M}}) = \frac{n}{2\sigma^2 m^2}\|\breve{M}\|_F^2 = \frac{n}{2\sigma^2 m^2}k^2\left\lfloor\frac{m}{r}\right\rfloor rm$$

Therefore, to satisfy condition 2 in theorem 1.3, we have to pick k satisfying,

$$\frac{n}{2\sigma^2 m^2} k^2 \left\lfloor \frac{m}{r} \right\rfloor rm \leq \frac{1}{4} \log(2^{rm/8}) = \frac{\log(2)}{4} \frac{rm}{8} \tag{2}$$

Finally, note that if $\breve{M}$ and $\breve{M}'$ are matrices in $\breve{\mathcal{M}}$, then

$$\begin{aligned} \|\breve{M} - \breve{M}'\|_F^2 &\geq k^2 |\{i,j : \breve{M}(i,j) \neq \breve{M}'(i,j)\}| \\ &\geq k^2 \left\lfloor \frac{m}{r} \right\rfloor \frac{rm}{8} \geq c \left((\sigma \wedge a)^2 \frac{mr}{n} m^2 \right) \end{aligned}$$

where c is an absolute constant, and in the last inequality we picked a k smaller than a and satisfying (2). Having checked all the conditions in theorem 1.3, we conclude the result as stated. □

1.2.2 Upper bound for noisy low-rank matrix completion

In the noisy case, we cannot recover the objective matrix M_* perfectly and we are faced with the task of finding a computationally tractable estimator with a performance close to this theoretical lower bound. We begin with the following estimator

$$\breve{M} = \frac{m^2}{n} \sum_{k=1}^{m} y_k (e_{i_k} \otimes e_{j_k})$$

where $\{e_1, \ldots, e_m\}$ is the standard basis on $\mathbb{R}^{m \times m}$. Note that $\mathbb{E}\breve{M} = M_*$. Although this estimator performs poorly when we measure its error in Frobenius norm, we can bound its spectral norm error using bounds on the sum of random matrices.

Theorem 1.4. Assume that $\eta_1, \ldots, \eta_n$ are i.i.d gaussian random variables with variance σ^2 and that M_* belongs to $\mathcal{M}(a, r)$. For every $t > 0$, $t_m := (t + \log(2m)) \log(m)$, $n \geq m t_m$, the following bound holds with probability at least $1 - 3e^{-t}$,

$$\|\breve{M} - M_*\| \leq 6m^2 (\sigma \vee a) \sqrt{\frac{t_m}{mn}} \tag{3}$$

Proof. From the definition of $\breve{M}$ and y_k, we conclude

$$\begin{aligned}\|\breve{M} - M_*\| &= \left\| \frac{m^2}{n} \sum_{k=1}^{n} [M_*(i_k, j_k)(e_i \otimes e_j) + \eta_k (e_i \otimes e_j)] - M_* \right\| \\ &\leq m^2 \left\| \frac{1}{n} \sum_{k=1}^{n} \left[M_*(i_k, j_k)(e_{i_k} \otimes e_{j_k}) - \frac{1}{m^2} M_* \right] \right\| \\ +m^2 \left\| \frac{1}{n} \sum_{k=1}^{n} \eta_k [(e_{i_k} \otimes e_{j_k}) - \mathbb{E}(e_{i_k} \otimes e_{j_k})] \right\| &+ m^2 \left\| \frac{1}{n} \sum_{k=1}^{n} \eta_k \mathbb{E}(e_{i_k} \otimes e_{j_k}) \right\| \\ &= m^2 \|\Xi_1\| + m^2 \|\Xi_2\| + m^2 \|\Xi_3\| \leq 3m^2 [\|\Xi_1\| \vee \|\Xi_2\| \vee \|\Xi_3\|]\end{aligned}$$

where

$$\Xi_1 := \frac{1}{n} \sum_{k=1}^{n} \left[M_*(i_k, j_k)(e_{i_k} \otimes e_{j_k}) - \frac{1}{m^2} M_* \right], \qquad \Xi_2 := \frac{1}{n} \sum_{k=1}^{n} \eta_k (e_{i_k} \otimes e_{j_k}).$$

$$\Xi_3 := \frac{1}{n} \sum_{k=1}^{n} \eta_k \mathbb{E}(e_{i_k} \otimes e_{j_k})$$

We proceed to bound $\|\Xi_1\|$, $\|\Xi_2\|$ and $\|\Xi_3\|$. We bound $\|\Xi_1\|$ using the following bounds on the spectral norm of sums of bounded random matrices [58]:

Lemma 1.5 (Non-commutative Bernstein inequality with bounded entries)**.** Let $Z_1, \ldots, Z_n$ be i.i.d $m \times m$ random matrices with $\mathbb{E}Z_k = 0$, $\sigma_Z^2 := \|\mathbb{E}Z_k^T Z_k\|$ and $\|Z_k\| \leq U$ for some $U > 0$. Then for all $t > 0$, with probability at least $1 - e^t$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\| \leq 2 \left(\sigma_Z \sqrt{\frac{t + \log(2m)}{n}} \bigvee U \frac{t + \log(2m)}{n} \right)$$

For the random matrices $Z_k := M_*(i_k, j_k)(e_{i_k} \otimes e_{j_k}) - \frac{1}{m^2} M_*$, $i = 1, \ldots, n$, we can easily check $\mathbb{E}Z_k = 0$, $\|\mathbb{E}Z_k^T Z_k\| \leq 2a^2/m$, and $\|Z_k\| \leq a$; and therefore, by the lemma we conclude that, for each $t > 0$, with probability at least $1 - e^t$,

$$\|\Xi_1\| \leq 2a \left(\sqrt{\frac{2(t + \log(2m))}{mn}} \bigvee \frac{t + \log(2m)}{n} \right)$$

We bound $\|\Xi_2\|$ using the following bound on the spectral norm of sum of sub-exponential random matrices,

Lemma 1.6 (Non-commutative Bernstain inequality with bounded moments)**.** Let $W_1, \ldots, W_n$ be i.i.d $m \times m$ random matrices with $\mathbb{E}W_k = 0$ and $\sigma_W^2 := \|\mathbb{E}W_k^T W_k\|$.

Suppose that,

$$\varphi_\alpha(W_k) := \inf\{u > 0 : \mathbb{E}\exp(\|W_k\|^\alpha)/u^\alpha \leq 2\} \leq U_\alpha$$

for some $\alpha \geq 1$. Then for all $t > 0$, with probability at least $1 - e^t$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i\right\| \leq 2\left(\sigma_W\sqrt{\frac{t+\log(2m)}{n}} \bigvee U_\alpha\left(\log\frac{U_\alpha}{\sigma_W}\right)^{1/\alpha}\frac{t+\log(2m)}{n}\right)$$

We apply this result to the random matrices $W_k := \eta_k[(e_{i_k} \otimes e_{j_k}) - \mathbb{E}(e_{i_k} \otimes e_{j_k})]$. A standard calculation shows that $\mathbb{E}W_k = 0$ and that $\|\mathbb{E}W_k^T W_k\| \leq 2\sigma^2/m =: \sigma_W$. Moreover, note that $\|W_k\| \leq |\eta_k|$, and thus $\varphi_2(W_k) \leq \sigma =: U_2$. Applying the lemma, we conclude that, for each $t > 0$, with probability at least $1 - e^{-t}$,

$$\|\Xi_2\| \leq 2\sigma\left(\sqrt{\frac{2(t+\log(2m))}{nm}} \bigvee \frac{(t+\log(2m))\sqrt{\log(m)}}{n}\right)$$

We wound Ξ_3 by noticing that,

$$\|\Xi_3\| = \left\|\frac{1}{n}\sum_{k=1}^{n}\eta_k\mathbb{E}(e_{i_k} \otimes e_{j_k})\right\| \leq \sqrt{\frac{1}{m}}\left|\frac{1}{n}\sum_{k=1}^{n}\eta_k\right| \tag{4}$$

and since η_k are zero mean normally distributed random variables with parameter σ, we conclude that, for every $t > 0$, with probability at least $1 - e^{-t}$, $\|\Xi_3\| \leq \sqrt{\frac{t}{nm}}$.

The result follows by combining the probabilistic bounds on $\|\Xi_1\|$, $\|\Xi_2\|$ and $\|\Xi_3\|$ using the union bound, and by simplifying the resulting expression by taking into account that $n \geq mt_m$. □

Our goal is to find an estimator with Frobenius norm error comparable to the bound given by theorem 1.2. To achieve that, we construct a second estimator based on either hard or soft thresholding of the singular values of the unbiased estimator $\breve{M}$ [15, 19, 35]. Let M be a matrix with singular value decomposition $M = \sum_{k=1}^{m}\sigma_k(u_k \otimes v_k)$, and let $\sigma_* > 0$ be a constant. We consider the matrices $M^h_{\sigma_*}$ and $M^s_{\sigma_*}$ obtained by hard and soft truncation of the singular values of M with threshold σ_*:

$$\begin{aligned} M^h_{\sigma_*} &= \sum_{\{k:\sigma_k > \sigma_*\}}\sigma_k(u_k \otimes v_k), \\ M^s_{\sigma_*} &= \sum_{\{k:\sigma_k > \sigma_*\}}(\sigma_k - \sigma_*)(u_k \otimes v_k). \end{aligned} \tag{5}$$

As a consequence of the celebrated Young-Eckart theorem [20], the hard-thresholding matrix $M^h_{\sigma_*}$ is a solution of the following rank minimization problem,

$$M^h_{\sigma_*} = \operatorname*{argmin}_{N\in\mathbb{C}^{m_1\times m_2}} \frac{1}{2}\|M-N\|_F^2 + \sigma_* \operatorname{rank}(N)$$

Similarly, the soft-thresholding matrix $M^s_{\sigma_*}$ is the solution of the following nuclear norm minimization problem (For a proof, see lemma 1.14)

$$M^s_{\sigma_*} = \operatorname*{argmin}_{N\in\mathbb{C}^{m_1\times m_2}} \frac{1}{2}\|M-N\|_F^2 + \sigma_* \|N\|_*$$

These optimization problems suggest that truncation of singular values might be used to solve low-rank matrix completion. In fact, using matrix perturbation theory, we study the *hard-thresholding estimator* $\breve{M}^h_{\sigma_*}$ and the *soft-thresholding estimator* $\breve{M}^s_{\sigma_*}$. First, we prove an upper bound for the Frobenius norm error depending on the nuclear norm of the target matrix M_*. Since the dependance on the number of samples is of the order $1/\sqrt{n}$, we refer to this kind of bound as a slow rate bound. To be precise, let us consider the following theorem.

Theorem 1.7. For each $t > 0$, $t_m := (t + \log(2m))\log(m)$, $n \geq mt_m$, and $\sigma_* := 2.02(\sigma \vee a)^2 m^2 \sqrt{\frac{t_m}{nm}}$. The following bound holds, with probability at least $1 - 3e^{-t}$

$$\frac{1}{m^2}\|\breve{M}_{\sigma_*} - M_*\|_F^2 \leq C(\sigma \vee a)\sqrt{\frac{t_m}{nm}}\|M_*\|_*$$

where $\breve{M}_{\sigma_*}$ is either the hard-thresholding estimator $\breve{M}^h_{\sigma_*}$ or the soft-thresholding estimator $\breve{M}^s_{\sigma_*}$.

Proof. The core of the theorem lies on the following perturbation theory inequality (for a proof of this perturbation inequality, see theorem 8.1 in [15]),

Lemma 1.8. For any pair of $m \times m$ matrices A and B, for $\sigma_* := (1+\delta)\|A-B\|$, and for any $\delta > 0.004$, there is an universal constant C such that,

$$\|A_{\sigma_*} - B\|_F^2 \leq C(1+\delta)\|A-B\|\|B\|_*$$

where A_{σ_*} is either the hard thresholding matrix $A^h_{\sigma_*}$ or the soft thresholding matrix $A^s_{\sigma_*}$.

Define δ by the following relation

$$(1+\delta)\|\breve{M}_{\sigma_*} - M_*\| = 6.06m^2(\sigma \vee a)\sqrt{\frac{t_m}{mn}}$$

Note that, when $n \geq mt_m$, theorem 1.4 implies $\|\breve{M} - M_*\| \leq 6m^2(\sigma \vee a)\sqrt{\frac{t_m}{mn}}$ with probability at least $1 - 3e^{-t}$. Thus,

$$(1+\delta) = \frac{6.06m^2(\sigma \vee a)\sqrt{\frac{t_m}{mn}}}{\|\breve{M}_{\sigma_*} - M_*\|} \geq \frac{6.06m^2(\sigma \vee a)\sqrt{\frac{t_m}{mn}}}{6m^2(\sigma \vee a)\sqrt{\frac{t_m}{mn}}} = 0.01$$

Therefore $\delta \geq 0.004$, and by lemma 1.8 with $A = \breve{M}$ and $B = M_*$, we conclude,

$$\begin{aligned}\|\breve{M}_{\sigma_*} - M_*\|_F^2 &\leq C(1+\delta)\|\breve{M} - M_*\|\|M_*\|_* \\ &= 6.06Cm^2(\sigma \vee a)\sqrt{\frac{t_m}{mn}}\|M_*\|_*\end{aligned}$$

□

Similarly, we prove an upper bound on the Frobenius norm error depending on the rank of the target matrix M_*. This upper bound matches the lower bound given by theorem 1.2. Since the dependance on the number of samples is of the order $1/n$, we refer to this type of bound as a fast rate bound. To be precise, let us consider the following theorem.

Theorem 1.9. There is an absolute constant C, such that for each $t > 0$, $t_m := (t + \log(2m))\log(m)$, $n \geq mt_m$, and $\sigma_* := 6m^2(\sigma \vee a)\sqrt{\frac{t_m}{nm}}$; the following bound holds, with probability at least $1 - 3e^{-t}$

$$\frac{1}{m^2}\|\breve{M}_{\sigma_*} - M_*\|_F^2 \leq C(\sigma \vee a)^2\frac{mt_m}{n}\text{rank}(M_*)$$

where $\breve{M}_{\sigma_*}$ is either the hard-thresholding estimator $\breve{M}^h_{\sigma_*}$ or the soft-thresholding estimator $\breve{M}^s_{\sigma_*}$.

Proof. Note that,

$$\begin{aligned}
\|\breve{M}_{\sigma_*} - M_*\|_F^2 &\leq \|\breve{M}_{\sigma_*} - M_*\|^2 \operatorname{rank}(\breve{M}_{\sigma_*} - M_*) \\
&\leq (\|\breve{M}_{\sigma_*} - \breve{M}\| + \|\breve{M} - M_*\|)^2 \operatorname{rank}(\breve{M}_{\sigma_*} - M_*) \\
&\leq \left(12m^2(\sigma \vee a)\sqrt{\frac{t_m}{nm}}\right)^2 \operatorname{rank}(\breve{M}_{\sigma_*} - M_*) \\
&= Cm^2(\sigma \vee a)^2 \frac{mt_m}{n} \operatorname{rank}(\breve{M}_{\sigma_*} - M_*)
\end{aligned}$$

where the last inequality holds with probability $1 - 3e^{-t}$ by theorem 1.4 and by the definition of $\breve{M}$. To bound the rank of $\breve{M}_{\sigma_*} - M_*$, we rely on the following classical lemma in perturbation theory (The proof of the lemma follows from Lidskii's theorem [8]),

Lemma 1.10. For any pair of $m \times m$ matrices A and B, the following inequality holds,

$$\max_{k=1,\ldots,m} |\sigma_k(A) - \sigma_k(B)| \leq \|A - B\|$$

where $\sigma_k(A)$ and $\sigma_k(B)$ are the singular values of A and B respectively in non increasing order.

By this lemma, if $\sigma_k(M_*) = 0$, then $|\sigma_k(\breve{M})| \leq \|\breve{M} - M_*\| \leq 6m^2(\sigma \vee a)\sqrt{\frac{t_m}{mn}}$, which implies $\sigma_k(\breve{M}_{\sigma_*}) = 0$. As a consequence, $\operatorname{rank}(\breve{M}_{\sigma}) \leq \operatorname{rank}(M_*)$. Therefore, $\operatorname{rank}(\breve{M}_{\sigma_*} - M_*) \leq \operatorname{rank}(\breve{M}_{\sigma_*}) + \operatorname{rank}(M_*) \leq 2\operatorname{rank}(M_*)$, and the result follows. □

1.3 Trace regression model and matrix LASSO

The study of matrix completion problems has been heavily influenced by the research on compressed sensing and sparse recovery. Let us consider the problem of recovering a vector $s_* \in \mathbb{R}^m$ based on random observations $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_k \in \mathbb{R}^m$, $k = 1, \ldots, m$, are random measurement vectors, and y_k, $k = 1, \ldots, n$ are random variables satisfying $\mathbb{E}(y_k|x_k) = x_k^T s_*$. We are interested in the case where the target vector s_* is sparse or it can be well approximated by a sparse vector. The support

of a vector s is the set of coordinates where s is different from zero. We measure the sparseness of a vector s by its ℓ_0-"norm", defined as the size of its support $\|s\|_{\ell_0} := |\{k \in [m] : s(k) \neq 0\}|$. We define the ℓ_1-norm of s as $\|s\|_{\ell_1} := \sum_{k=1}^{m} |s(k)|$. We think of the ℓ_1-norm as a convex approximation of the ℓ_0-"norm".

Sparse recovery deals with this problem in the case where we cannot alter the design of the measurement vectors x_k ; while compressed sensing consider the case where we can design the distribution of the measurement vectors x_k. The *least absolute shrinkage and selection operator (LASSO)* [57], a classic estimator for sparse recovery, estimates s_* as the solution of the following regularized least-squares problem:

$$\hat{s} = \underset{s \in \mathbb{R}^m}{\operatorname{argmin}} \sum_{k=1}^{n} (y_k - x_k^T s)^2 + \varepsilon \|s\|_{\ell_1} \tag{6}$$

where $\varepsilon > 0$ is a regularization parameter. In this optimization procedure, we would prefer to penalize using the ℓ_0-norm, that is the number of nonzero entries of vector s. Similarly to rank minimization, this problem is NP-hard, and therefore we relax it to an ℓ_1-norm minimization problem.

1.3.1 Trace regression model

In this section we consider a generalization of this estimation procedure for matrices. Let us consider the problem of estimating an $m \times m$ matrix M_* based on observations $(X_1, y_1), \ldots, (X_n, y_n)$ where X_k, $k = 1, \ldots, n$ is an $m \times m$ random matrix with distribution Π_k, and y_k, $k = 1, \ldots, n$ is a random variable satisfying the *trace regression model*, that is,

$$\mathbb{E}(y_k | X_k) = \langle M_*, X_k \rangle, \qquad k = 1, \ldots, n$$

In this context, we refer to the matrices X_k, $k = 1, \ldots, n$ as *design matrices* and to the observations Y_k, $k = 1, \ldots, n$ as the *response variables*. It is often convenient to express the response variables as $y_k = \langle M_*, X_k \rangle + \eta_k$, $k = 1, \ldots, n$, where $\eta_k := y_k - \mathbb{E}(y_k | X_k)$ are zero mean random variables representing noise.

Let Π be the average distribution of the design matrices, that is, $\Pi = \frac{1}{n}\sum_{k=1}^{n}\Pi_k$, we introduce the design dependent inner product and its induced norm,

$$\langle M, N\rangle_{L_2(\Pi)} := \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\langle M, X_k\rangle\langle N, X_k\rangle, \qquad \|M\|^2_{L_2(\Pi)} := \langle M, M\rangle_{L_2(\Pi)}$$

As exemplified below; in the trace regression framework, we can model matrix completion, point masks, complete subgaussian designs and fixed design among other matrix estimation problems.

Example (Matrix Completion)**.** We recover the *matrix completion* scenario when the design matrices X_i are i.i.d copies from a random matrix X with distribution Π on the set

$$\mathcal{X} = \{e_i \otimes e_j \in \mathbb{C}^{m\times m} : 1 \le i, j \le m\}.$$

where e_k, $k = 1, \dots, m$ are the vectors of the canonical basis in $\mathbb{R}^m$. When Π is the uniform distribution, we recover the widely study *uniform sampling matrix completion* scenario. It is possible to consider even more general matrix measurement models in which, for a given orthonormal basis in the space of matrices, a random sample of Fourier coefficients of the target matrix M_* is observed subject to a random noise.

Example (Collaborative sampling)**.** As in matrix completion, in *collaborative sampling*, the design matrices X_i are sampled from the set

$$\mathcal{X} = \{e_i \otimes e_j \in \mathbb{C}^{m\times m} : 1 \le i, j \le m\}.$$

but the each sampled matrix is different than the previous one. Therefore, the distributions Π_k, $k = 1, \dots, n$ are not independent.

Example (Point masks)**.** Instead of sampling from only one entry of the matrix, we can consider the case where we observe averages of a group of entries. To be precise, consider the case where the design matrices are sampled from the set

$$\mathcal{X} = \left\{\sum_{k=1}^{K} e_i \otimes e_j \in \mathbb{C}^{m\times m} : 1 \le i, j \le m\right\}$$

where K is a typically small number. Clearly, when $K = 1$, this case reduces to the matrix completion case.

Example (Column masks)**.** In the *column mask scenario*, we consider the design matrices X_k as i.i.d. copies of a random matrix X, which has only one nonzero column. For instance, let the distribution of X be such that all the columns have equal probability to be non-zero, and the random entries of non-zero column $x(j)$ are such that $\mathbb{E}(x(j)x(j)^T)$ is the identity matrix. In multitask learning, one can be interested in considering non-identically distributed X_k. The model can be then reformulated as a longitudinal regression model, with different distributions of X_k corresponding to different tasks [55].

Example ("Complete" subgaussian design)**.** In the *complete subgaussian scenario*, we assume that the design matrices X_k are i.i.d. copies of a random matrix X such that $\langle M, X\rangle$ is a subgaussian random variable for any matrix M. This approach has its roots in compressed sensing. The two major examples are given by the matrices X whose entries are either i.i.d. standard Gaussian or Rademacher random variables.

Example (Fixed design)**.** We can model the case of non-random design matrices by setting all the Π_k, $k = 1, \dots, n$ as Dirac measures. In particular, when M_* and X_k, $k = 1, \dots, n$, are diagonal matrices, the trace regression model becomes the usual linear regression model. In that case, the rank of M_* becomes the number of its non-zero diagonal elements. This observation allows us to study the usual LASSO in sparse linear regression with fixed design.

1.3.2 Matrix LASSO

In the case where matrix M_* is low rank or it can be well approximated by a low rank matrix, we consider the following matrix LASSO estimator

$$\hat{M} := \hat{M}_{(\varepsilon;\mathcal{M})} := \operatorname*{argmin}_{M\in\mathcal{M}} \frac{1}{n}\sum_{k=1}^{n}(y_k - \langle X_k, M\rangle)^2 + \varepsilon\|M\|_* \tag{7}$$

where $\varepsilon > 0$ is a regularization parameter and $\mathcal{M}$ is a convex domain in the space of $m \times m$ matrices. The matrix LASSO estimator has been studied by several authors under different conditions on the target and design matrices. In [45], Ma, Goldfarb and Chen introduce the matrix LASSO and develop algorithms to solve it efficiently using fixed point continuation. In [12], Candes and Plan derive oracle inequalities for the matrix LASSO using a matrix version of the restricted isometry conditions used in the analysis of standard LASSO. In [55], Rohde and Tsybakov develop non-asymptotic upper bounds for a general version of matrix LASSO where the regularization term is given by Shatten-p norms. In [47], Negahban and Wainwright analyze the matrix LASSO for the case where the target matrix is low-rank and non-spiky. In [40], Koltchinskii, Lounici and Tsybakov exploit the knowledge of the design distributions to derive lower and upper bounds for a linearized version of the matrix LASSO. In [39], Koltchinskii develops tight oracle inequalities for a general version of the matrix LASSO with a quadratic-type loss function. In [36], Klopp derives upper bounds for the case where the optimization domain is the set of matrices with bounded entries.

As an example of the performance of matrix LASSO, let us consider the case of matrix completion under uniform sampling where the response variables are contaminated by zero mean gaussian noise with variance σ^2. Let us assume that the target matrix M_* belongs to the set $\mathcal{M}(a, r)$ of $m \times m$ matrices of rank r and entries bounded by a constant a. The following theorem, proved by Klopp in [36], shows that with a proper choice of the regularization parameter ε, and a proper choice of the optimization domain, matrix LASSO achieves optimal rates up to log factors.

Theorem 1.11 (Klopp [36])**.** For $t > 0$, $t_m := t + \log(2m)$, $n \geq m t_m$ and $\varepsilon = 4(\sigma \vee a)\sqrt{\frac{t_m}{mn}}$, the following bound holds for the estimator $\hat{M}_{(\varepsilon;\mathcal{M}_a)} = \hat{M}$, with probability $1 - e^{-t}$,

$$\frac{1}{m^2}\|\hat{M} - M_*\|_F^2 \leq C(a \vee \sigma)^2 \frac{m t_m}{n} \mathrm{rank}(M_*)$$

where $\mathcal{M}_a$ is the convex set of $m \times m$ matrices with entries bounded by a, and C is

an absolute constant.

1.3.3 Linearized matrix LASSO

In [40], Koltchinskii, Lounici and Tsybakov present the following linearized version of matrix LASSO for the case where the design distributions Π_k, $k = 1, \ldots, n$ are known,

$$\tilde{M}_\varepsilon := \tilde{M} := \operatorname*{argmin}_{M \in \mathcal{M}} \|M\|^2_{L_2(\Pi)} + \frac{2}{n} \sum_{k=1}^{n} \langle y_k X_k, M \rangle + \varepsilon \|M\|_* \tag{8}$$

where $\Pi = \frac{1}{n} \sum_{k=1}^{n} \Pi_k$, and $\varepsilon \geq 0$ is a regularization parameter. We define the stochastic matrix Ξ as follows,

$$\Xi = \frac{1}{n} \sum_{k=1}^{n} (y_k X_k - \mathbb{E}(y_k X_k))$$

The following oracle inequality holds under the assumptions that $\mathcal{M}$ is a convex set, and that there is a constant $\mu > 0$ such that $\|M\|^2_{L_2(\Pi)} \geq \mu^{-2} \|M\|_F$, for each $M \in \mathcal{M} - \mathcal{M} := \{M_1 - M_2 : M_1 \in \mathcal{M}, M_2 \in \mathcal{M}\}$,

Theorem 1.12 (Koltchinskii, Lounici, Tsybakov [40])**.** If $\varepsilon \geq 2\|\Xi\|$, then

$$\|\tilde{M} - M_*\|_{L_2(\Pi)} \leq \inf_{M \in \mathcal{M}} \left[\|M - M_*\|_{L_2(\Pi)} + C\mu^2 \varepsilon^2 \operatorname{rank}(M) \right]$$

where C is an absolute constant.

In the case of matrix completion under uniform sampling, the design matrix X_k is equal to $e_{i_k} \otimes e_{j_k}$, where i_k and j_k are indexes chosen independently uniformly at random from $\{1, \ldots, m\}$. As a consequence, $\|M\|^2_{L_2(\Pi)} = m^{-2} \|M\|^2_F$, and the linearized matrix LASSO estimator reduces to

$$\begin{aligned}
\tilde{M}_\varepsilon &= \operatorname*{argmin}_{M \in \mathcal{M}} \frac{1}{m^2} \left[\|M\|^2_F + 2 \left\langle \frac{m^2}{n} \sum_{k=1}^{n} y_k (e_{i_k} \otimes e_{j_k}), M \right\rangle \right] + \varepsilon \|M\|_* \\
&= \operatorname*{argmin}_{M \in \mathcal{M}} \left[\|M\|^2_F + 2 \left\langle \breve{M}, M \right\rangle \right] + m^2 \varepsilon \|M\|_* \\
&= \operatorname*{argmin}_{M \in \mathcal{M}} \|M - \breve{M}\|^2_F + m^2 \varepsilon \|M\|_*
\end{aligned}$$

where $\breve{M}$ is the unbiased estimator introduced in section 1.2.2. The solution to this optimization problem is the soft-thresholding estimator $\breve{M}^s_{2m^2\varepsilon}$.

Note that, in this case, the stochastic matrix Ξ is equal to $\frac{1}{m^2}(\breve{M} - M)$. When the response variables are contaminated by zero mean gaussian noise with variance σ^2 and the target matrix has entries bounded by a, for each $t > 0$, $t_m := (t + \log(2m))\log(m)$ and $n > mt_m$, we obtain the following bound on the stochastic error $\|\Xi\|$ using bounds on the spectral norm of the sum of random matrices (compare to theorem 1.4),

$$\|\Xi\| \leq 6(\sigma \vee a)\sqrt{\frac{t_m}{mn}}$$

We can apply theorem 1.12 to derive an optimal (up to log factors) oracle inequality for the linearized LASSO estimator,

Theorem 1.13. For $t > 0$, $t_m = (t + \log(2m))\log(m)$, $n \geq mt_m$, $\varepsilon = 12(\sigma \vee a)\sqrt{\frac{t_m}{mn}}$ and an arbitrary matrix M, the following bound holds for the estimator $\tilde{M}_\varepsilon = \tilde{M}$, with probability $1 - e^{-t}$,

$$\frac{1}{m^2}\|\tilde{M} - M_*\|_F^2 \leq \frac{1}{m^2}\|M - M_*\|_F^2 + C(a \vee \sigma)^2 \frac{mt_m}{n}\operatorname{rank}(M)$$

where C is an absolute constant.

1.4 Proximal Algorithms

The LASSO, the matrix LASSO, and the linearized matrix LASSO estimators are well-structured convex optimization problems that can be solved in a theoretically efficient fashion by using polynomial-time interior point methods [37, 42, 24, 44]. Nevertheless, the time complexity of interior point methods has cubic dependence on the dimension of the problem. Since most applications lead to extremely large-scale problems, this cubic dependance makes interior point methods impractical. In contrast, we consider proximal methods that, when properly designed, lead to nearly dimension-independent rates of convergence [51, 49, 43, 3]. The main disadvantage of

proximal methods is that their rate of convergence is only sub-linear with inaccuracy tending to zero with the number of iterations k at a rate of $O(1/k^2)$ at best, or even $O(1/k)$. However, in the majority of applications of nuclear norm minimization, we are only interested in medium-accuracy solutions, and therefore, the relatively slow convergence of proximal methods is compensated by the insensitivity to problem size.

1.4.1 Proximal operator

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a *proper closed convex function*; that is a function f with an epigraph $\text{epi}(f) := \{(x, y) \in \mathbb{R}^d \times \mathbb{R} : f(x) \leq t\}$ is a nonempty closed convex set. We define the *proximal operator* $\text{prox}_f : \mathbb{R}^d \to \mathbb{R}$ of a proper closed convex function f by

$$\text{prox}_f(x) = \underset{z \in \mathbb{R}^d}{\text{argmin}} \left[f(z) + \frac{1}{2}\|z - x\|_2^2 \right] \tag{9}$$

where $\|\cdot\|_2$ is the standard euclidean norm of $\mathbb{R}^d$. Note that proximity operator is well defined, since the objective function in (10) is a proper strictly convex function, and therefore there is a unique minimizer for every $x \in \mathbb{R}^d$. For a parameter $\lambda > 0$, we often consider the proximal operator of the scaled function λf, which can be expressed as

$$\text{prox}_{\lambda f}(x) = \underset{z \in \mathbb{R}^d}{\text{argmin}} \left[f(x) + \frac{1}{2\lambda}\|z - x\|_2^2 \right] \tag{10}$$

An useful property of the proximity operator is that a point $x_* \in \mathbb{R}^d$ minimizes a proper strictly convex function $f : \mathbb{R}^d \to \mathbb{R}$ if and only if $x_* = \text{prox}_f(x_*)$. Therefore, one can minimize f by finding fixed points of prox_f. A convex function f is called *strongly convex* when for all $x_1, x_2 \in \mathbb{R}^d$, the following inequality holds for each $t \in (0, 1)$

$$f(tx_1 + (1-t)x_2) \leq tf(t) + (1-t)f(t) - \frac{1}{2}t(1-t)\|x_1 - x_2\|_2^2$$

When f is strongly convex, the operator prox_f turns out to be a *contraction*, that is a Lipschitz continuous operator with constant less than 1, and therefore repeatedly

applying prox_f finds a fixed point. For general convex functions, prox_f is not necessarily a contraction, but it always is a *firm non-expansive operator*; that is, prox_f satisfies the following inequality for all $x_1, x_2 \in \mathbb{R}^d$

$$\| \text{prox}_f(x_1) - \text{prox}_f(x_2)\|_2^2 \leq (x_1 - x_2)^T(\text{prox}_f(x_1) - \text{prox}_f(x_2))$$

Firm non expansive operators are sufficient for fixed point iterations. Thus, the so-called *proximal point algorithm*, defined by the following iterative procedure, will converge whenever a minimizer exists,

$$x_{k+1} := \text{prox}_{\lambda f}\, x_k$$

1.4.2 Proximal gradient method

We are interested in solving optimization problems of the form

$$x_* = \underset{x \in \mathbb{R}^d}{\text{argmin}}\, f(x) + g(x) \tag{11}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ are closed proper convex functions, and moreover f is differentiable. Note that we can use g to encode constrains, since it takes values on the extended real line. The *proximal gradient method* uses the following iteration,

$$x_{k+1} := \text{prox}_{\lambda_k g}(x_k - \lambda_k \nabla f(x_k))$$

where λ_k is a step size. For this procedure, we can guarantee a rate of convergence of $O(1/k)$, when the step size is chosen as a constant $\lambda_k = \lambda \in (0, 2L]$, where L is the Lipschitz constant of ∇f. In practical scenarios, the step size is found in each step by line search. Table 1 shows a pseudocode for the proximal method where the parameter λ is chosen iteratively using a backtracking algorithm proposed in [6].

1.4.3 Accelerated proximal gradient method

By including an extrapolation step in the iteration of the proximal gradient method, we are able to improve its rate of convergence. For a simple version of this idea, let

Table 1: Proximal algorithm with backtracking

Inputs: A convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, a convex function $g : \mathbb{R}^d \to \mathbb{R}$, an initial point x_0, a initial step size λ_0, and a line search parameter $\beta \in (0, 1)$.

Initialize $k := 0$ and $\lambda := \lambda_0$.

Repeat until a stoping criteria is satisfied.

Set $z := \text{prox}_{\lambda g}(x_k - \lambda \nabla f(x_k))$.
If $f(z) + \nabla f(x_k)^T(x_k - z) < f(x_k) + \frac{1}{\lambda}\|x_k - z\|_2^2$.
Update $k \leftarrow k + 1$ and $\lambda \leftarrow \lambda_0$.
Set $x_k := z$.
Update $\lambda \leftarrow \beta\lambda$.

Output: A near optimal solution x_k to (11) satisfying $\|x_k - x_*\|_2^2 = O(1/k)$.

us consider the following iteration,

$$y_{k+1} := x_k + \omega_k(x_k - x_{k-1})$$

$$x_{k+1} := \text{prox}_{\lambda_k g}(y_{k+1} - \lambda_k \nabla f(y_{k+1}))$$

where ω_k is an extrapolation, and $\lambda_k \in [0, 1)$ is the usual step size. The extrapolation parameter can be chosen as $\omega_k := k/(k + 3)$, while the step size can be chosen as a constant $\lambda_k = \lambda \in (0, L]$, where L is the Lipschitz constant of ∇f. In practical scenarios, the step size can be found in each step by line search. By choosing a proper choice of the parameters, we can achieve an "accelerated" rate of convergence of $O(1/k^2)$. Nesterov coined the term *accelerated first order method* because it has a worst-case convergence rate that is superior to the standard methods and that cannot be improved further [48]. Table 2 shows a pseudocode for the accelerated proximal algorithm where the step size is picked via sidetracking [6].

Table 2: Accelerated proximal algorithm with backtracking

Inputs: A convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, a convex function $g : \mathbb{R}^d \to \mathbb{R}$, an initial point x_0, an initial step size λ_0, and a line search parameter $\beta \in (0, 1)$.

Initialize $k := 0$, $\lambda := \lambda_0$ and $y_0 := x_0$.

Repeat until a stoping criteria is satisfied.

Set $\omega_k := \frac{k}{k+3}$.

Set $y_k := x_k + \omega_k(x_k - x_{k-1})$.

Set $z := \text{prox}_{\lambda g}(y_k - \lambda \nabla f(y_k))$.

If $f(z) + \nabla f(y_k)^T(y_k - z) < f(y_k) + \frac{1}{\lambda}\|y_k - z\|_2^2$.

Update $k \leftarrow k + 1$ and $\lambda \leftarrow \lambda_0$

Set $x_k := z$.

Update $\lambda \leftarrow \beta\lambda$.

Output: A near optimal solution x_k to (11) satisfying $\|x_k - x_*\|_2^2 = O(1/k^2)$.

1.4.4 An accelerated proximal gradient algorithm for matrix LASSO

The matrix LASSO estimator falls into the category of non-smooth convex optimization problems that we can solve by proximal gradient and accelerated proximal gradient algorithms. We consider the space of $m \times m$ real valued matrices as the euclidean space $\mathbb{R}^{m\times m}$, and we choose $f : \mathbb{R}^{m\times m} \to \mathbb{R}$ and $g : \mathbb{R}^{m\times m} \to \mathbb{R}$ in (11) as follows:

$$f(M) := \frac{1}{n}\sum_{k=1}^{n}(y_k - \langle X_k, M\rangle)^2, \qquad g(M) := \varepsilon\|M\|_*.$$

The following lemma provides us with an efficient method to calculate the proximal operator of the nuclear norm.

Lemma 1.14. For any $m \times m$ real-valued matrix M, and every positive number τ, the following holds,

$$\text{prox}_{\tau\|\cdot\|_*}(M) := \underset{N\in M}{\text{argmin}}\, \frac{1}{2}\|N - M\|_F^2 + \tau\|N\|_* = M_\tau^s \tag{12}$$

where M_τ^s is the soft-thresholding matrix defined in (5).

Proof. We define the *sub-differential* $\partial f(x)$ of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ as the set,

$$\partial f(x) := \{z \in \mathbb{R}^d : z^T(y - x) \leq f(y) - f(x), \forall y \in \mathbb{R}^d\}$$

A *sub-gradient* of f at x is a vector $z \in \partial f(x)$. A vector $\hat{x} \in \mathbb{R}^d$ minimizes f if and only if 0 is a sub-gradient of f at the vector $\hat{x}$. We proceed to prove that 0 is a sub gradient of the strictly convex function $h_{\tau,M} : N \in \mathbb{R}^{m\times m} \mapsto \frac{1}{2}\|N-M\|_F^2+\tau\|N\|_* \in \mathbb{R}$ at the matrix M_τ^s.

Let $M = \sum_{k=1}^m \sigma_k(u_k \otimes v_k)$ be the singular value decomposition of M and thus, by definition, $M_\tau^s = \sum_{\sigma_k>\tau}(\sigma_k - \tau)(u_k \otimes v_k)$. On one hand, any sub-gradient V of the nuclear norm at M_τ^s can be represented as (see [60])

$$V = \sum_{\sigma>\tau}(u_k \otimes v_k) + W$$

where $W \in \mathbb{R}^{m \times m}$ satisfies

$$\|W\| \leq 1, \qquad \sum_{\{j:\sigma_i > \tau\}} \sum_{\{i:\sigma_j > \tau\}} |\langle W, u_i \otimes v_j \rangle| = 0 \tag{13}$$

On the other hand, any sub-gradient of $h_{\tau,M}$ at M_τ^s can be represented as $M_\tau^s - M + \tau V$ where $V \in \partial \|M_\tau^s\|_*$. Therefore, if $\tau^{-1}(M - M_\tau^s)$ is a sub-gradient of the nuclear norm at the point M_τ^s, we would conclude that $0 \in \partial h_{\tau,M}(M_\tau^s)$, and the lemma will follows. Note that

$$\begin{aligned} \frac{1}{\tau}(M - M_\tau^s) &= \frac{1}{\tau}\left(\sum_{k=1}^{r} \sigma_k (u_k \otimes v_k) - \sum_{\sigma_k > \tau} (\sigma_k - \tau)(u_k \otimes v_k)\right) \\ &= \sum_{\sigma_k > \tau} (u_k \otimes v_k) + \sum_{\sigma_k \leq \tau} \frac{\sigma_k}{\tau}(u_k \otimes v_k) \end{aligned}$$

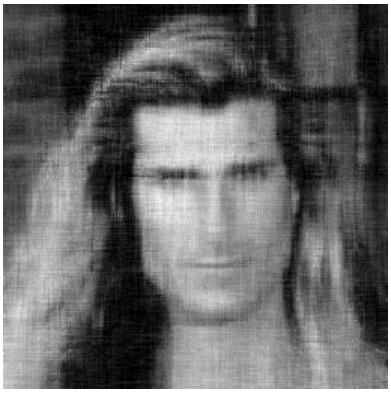
From standard algebraic calculations, we can check $W = \sum_{\sigma_k \leq \tau} \frac{\sigma_k}{\tau}(u_k \otimes v_k)$ satisfies the properties in (13), therefore $\tau^{-1}(M - M_\tau^s)$ is a sub-gradient of the nuclear norm at M_τ^s, and the result follows. □

Now that we have a procedure to calculate the proximal operator of the nuclear norm, we are in shape to solve the matrix LASSO optimization problem (7) using the accelerated proximal algorithm with backtracking shown in table 2. To exemplify this procedure, we revisit the problem of recovering a grayscale digital image from some observations of its pixels. As before, we represent each image by a matrix with entries between -1 and 1. Each observation consists of an index of the matrix chosen uniformly at random, and the value of that index contaminated with zero mean gaussian noise. The matrices representing the images are normalized to make their entries between -1 and 1. We consider the same two images studied in section 1.1.3. In figure 3, we present the recovered images using matrix LASSO under different levels of noise.

(a) Recovered Lenna. Standard deviation of noise $\sigma = 0.01$

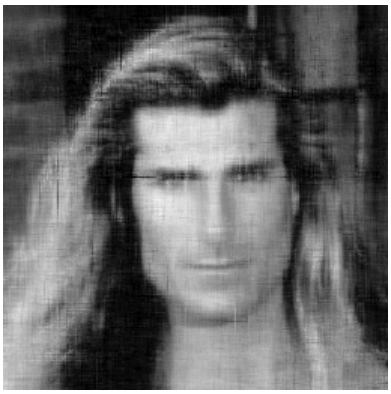
(b) Recovered Fabio. Standard deviation of noise $\sigma = 0.01$

(c) Recovered Lenna. Standard deviation of noise $\sigma = 0.005$



(d) Recovered Fabio. Standard deviation of noise $\sigma = 0.005$

(e) Recovered Lenna. Standard deviation of noise $\sigma = 0.001$



(f) Recovered Fabio. Standard deviation of noise $\sigma = 0.001$

Figure 3: Recovering Lenna and Fabio using matrix LASSO from 30.000 samples contaminated with gaussian noise with variance σ^2

Chapter II

LOW RANK ESTIMATION OF SIMILARITIES ON GRAPHS

With the hope of increasing its connectivity, a *social network site* [9] commissions us to develop a system for providing users with recommendations of people to invite into their circle of friends. A social network ability to proliferate depends strongly on its ability to provoke users to connect to each other, therefore, the problem of predicting potential friendships accurately is highly important for its survival [30]. To solve this problem, we base our strategy on predicting accurately the similarity between users. With that idea in mind, our goal is to design an estimator for similarities based on two kind of information: 1) The social network architecture, and 2) similarity information between some random pairs of members in the network.

2.1 Modeling the problem

We model the social network architecture by a *simple graph* $G = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a finite set of vertices representing users, and $\mathcal{E}$ is the set of edges representing links between users. Let $(U, V, Y) \in \mathcal{V} \times \mathcal{V} \times \{-1, +1\}$ be a random triple, where U and V are independent uniformly distributed vertices and Y is a label indicating the degree of *similarity* between U and V. More precisely, $Y = +1$ indicates that the vertices U and V are similar, while $Y = -1$ indicates that they are not. We refer to Y as the similarity between U and V. The conditional distribution of the similarity Y given U and V is completely characterized by the regression function

$$S_*(u, v) := \mathbb{E}(Y|U = u, V = v), \qquad u, v \in \mathcal{V}$$

where S_* is a real valued function on $\mathcal{V} \times \mathcal{V}$ such that $S_*(u, v) = S_*(v, u)$ for all $u, v \in \mathcal{V}$. In what follows, we refer to this kind of functions as *similarity kernels* over $\mathcal{V}$. We usually identify the linear space of similarity kernels with the space of real-valued symmetric matrices of dimension $|\mathcal{V}|$ and we denote it by $\mathcal{S}_\mathcal{V}$.

Our goal is to find a predictor g for the similarity Y based on U and V. Namely, a function $g : \mathcal{V} \times \mathcal{V} \to \{-1, 1\}$ able to predict the similarity between two vertices u and v correctly. We measure the performance of a predictor g by its *generalization error*

$$\mathbb{P}\{Y \neq g(U, V)\}$$

A *Bayes classifier* is a predictor that minimizes the generalization error. In the setting of our problem, the Bayes classifier is given by the function that maps each pair of vertices (u, v) to $\text{sign}(S_*(u, v))$. Therefore, the problem of finding a predictor for Y based on U and V can be reduced to the problem of estimating S_* as accurate as possible.

We base our estimate of S_* on training data $(U_1, V_1, Y_1), \ldots, (U_n, V_n, Y_n)$ consisting of n i.i.d. copies of (U, V, Y). We consider situations where S_* is a kernel of relatively small rank that possesses some degree of "smoothness" on the graph. On one hand, we justify the low-rank assumption by the belief that there are few underlying features characterizing the behavior of similarities. While, on the other hand, we justify the smoothness assumption by the belief that close vertices share some degree of similarity.

2.2 Characterizing smoothness

A *simple graph* G is a pair $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is an arbitrary set and $\mathcal{E}$ is a collection of 2-element subsets of $\mathcal{V}$. The elements of $\mathcal{V}$ are called vetices, and the elements of $\mathcal{E}$ are called edges. When a 2-element set $\{u, v\} \subseteq \mathcal{V}$ is an edge, we say that u and v are neighbors and we write $u \sim v$. The number of neighbors of a vertex u is called the

degree of u and it is denoted by $\deg(u)$. We identify the space of real-valued functions over $\mathcal{V}$ with the euclidean space $\mathbb{R}^{\mathcal{V}}$ endowed with the standard inner product $\langle \cdot, \cdot \rangle$ and the euclidean norm $\| \cdot \|$. Note that we are using the same notation $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ for the Hilbert-Schmidt inner product and for the spectral norm respectively. It is our hope that this little abuse of notation will cause no confusion to the attentive reader. We characterize the smoothness of a function $f : \mathcal{V} \to \mathbb{R}$ by its *energy*,

$$\mathscr{E}^2(f) = \sum_{u \sim v} |f(u) - f(v)|^2$$

In order to study the energy function $\mathscr{E}$, we introduce the *Laplacian* Δ_G of G,

$$\Delta_G(u,v) = \Delta(u,v) := \begin{cases} \deg(u) & u = v \\ -1 & u \sim v \\ 0 & u \not\sim v \end{cases}$$

In this context, we are interested in the Laplacian because it induces a positive semi-definite bilinear form. The induced seminorm calculates precisely the energy of functions as defined before. In other words, the Laplacian induces a geometry on the graph that is compatible with our measure of energy. To be precise,

$$\langle f, g \rangle_\Delta := \langle f, \Delta g \rangle = \langle \Delta^{1/2} f, \Delta^{1/2} g \rangle$$

$$\|f\|_\Delta^2 := \langle f, f \rangle_\Delta = \mathscr{E}^2(f)$$

We characterize the smoothness of a symmetric kernel $S : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ in terms of *Sobolev type norms* $\|\Delta^{p/2} S\|_F^2$ for some $p > 0$. Note that if S is a kernel of rank r with spectral representation $S = \sum_{k=1}^r \mu_k (\psi_k \otimes \psi_k)$, then

$$\|\Delta^{p/2} S\|_F^2 = \operatorname{tr}(\Delta^{p/2} S^2 \Delta^{p/2}) = \operatorname{tr}(\Delta^p S^2)$$
$$= \sum_{k=1}^r \mu_k^2 \langle \Delta^p \psi_k, \psi_k \rangle = \sum_{k=1}^r \mu_k^2 \|\Delta^{p/2} \psi_k\|^2,$$

so, essentially, the smoothness of the kernel S depends on the smoothness of its eigenfunctions ψ_k on the graph. In particular, for $p = 1$, we have

$$\|\Delta^{1/2} S\|_F^2 = \sum_{k=1}^r \mu_k^2 \sum_{u \sim v} |\psi_k(u) - \psi_k(v)|^2,$$

2.3 Estimation method

Without loss of generality, we assume that the vertex set of the graph $\mathcal{V}$ is the set of the first m positive natural numbers $[m] := \{1, \ldots, m\}$. We begin with the following estimator

$$\breve{S} = \frac{m^2}{n} \sum_{k=1}^{m} Y_k (e_{U_k} \otimes e_{V_k})$$

where $\{e_1, \ldots, e_m\}$ is the standard basis on $\mathbb{R}^m$. Note that $\mathbb{E}\breve{S} = S_*$. Although this estimator performs poorly when we measure its error in Frobenius norm, for several cases of interest, we can control its spectral norm error using bounds on the sum of random matrices.

Given a kernel S, let $L_n(S)$ denote the following penalized empirical risk:

$$L_n(S) := \frac{1}{m^2} \|S - \breve{S}\|_F^2 + \varepsilon_* \|S\|_* + \frac{\varepsilon_F}{m^2} \|W^{1/2} S\|_F^2 \tag{14}$$

where $W := \Delta^p$ for some constant $p > 0$, and for some regularization parameters ε_* and $\varepsilon_F > 0$. We analyze the following extremum estimator:

$$\hat{S} := \operatorname*{argmin}_{S \in \mathbb{S}} L_n(S), \tag{15}$$

where $\mathbb{S}$ is a closed convex subset of the linear space $\mathcal{S}_\mathcal{V}$ of all symmetric kernels. Note that there are two complexity penalties involved in the definition of penalized empirical risk (14). The first penalty is based on the nuclear norm $\|S\|_*$ and it is used to "promote" low rank solutions. The second penalty is based on a "Sobolev type norm" $\|W^{1/2} S\|_F^2$ and it is used to "promote" the smoothness of the solution on the graph. In principle, W in the definition of $L_n(S)$ could be an arbitrary symmetric positive semi-definite matrix. Therefore, alternative interpretations of the problem under consideration are possible. For instance, we can design a matrix W to learn similarities on weighted graphs or on Hilbert spaces.

Our goal is to derive an upper bound on the error $\|\hat{S} - S_*\|_F^2$ of estimator $\hat{S}$ in terms of spectral characteristics of the target similarity kernel S_* and the matrix W.

2.4 Spectral characteristics of S_* and W

2.4.1 Spectral properties of W

Suppose that W has the following spectral representation $W = \sum_{k=1}^{m} \lambda_k(\phi_k \otimes \phi_k)$, where $0 \le \lambda_1 \le \cdots \le \lambda_m$ are the eigenvalues of W (repeated with their multiplicities) and $\phi_1, \ldots, \phi_m$ are the corresponding orthonormal eigenfunctions (of course, there is a multiple choice of ϕ_k in the case of repeated eigenvalues). Let k_0 be the smallest k such that $\lambda_k > 0$. We will assume that for some (arbitrarily large) $\zeta \ge 1$, $\lambda_m \le m^{\zeta}$ and $\lambda_{k_0} \ge m^{-\zeta}$. In addition, it is assumed that, for some constant $c > 1$ and for all $k = k_0, \ldots, m-1$, $\lambda_{k+1} \le c\lambda_k$. The following spectral function characterizes the distribution of the eigenvalues:

$$F(\lambda; W) = F(\lambda) := \sum_{j=1}^{m} I(\lambda_j \le \lambda), \qquad \lambda \ge 0.$$

Our goal is to express our bounds in terms of spectral function F; nevertheless, due to some technicalities in the proof, we rely on an upper bound $\bar{F}(\lambda) \ge F(\lambda)$ that possesses some "regularity" in the sense that $\lambda \mapsto \frac{\bar{F}(\lambda)}{\lambda}$ is a nonincreasing function and, for some $\gamma \in (0, 1)$,

$$\int_{\lambda}^{\infty} \frac{\bar{F}(t)}{t^2} dt \le \frac{1}{\gamma} \frac{\bar{F}(\lambda)}{\lambda}, \qquad \lambda > 0.$$

It is easy to see that the last two conditions are satisfied if $\lambda \mapsto \frac{\bar{F}(\lambda)}{\lambda^{1-\gamma}}$ is a nonincreasing function and that the smallest upper bound on T with this property is

$$\bar{F}(\lambda) = \sup_{s \le \lambda} s^{1-\gamma} \sup_{t \ge s} \frac{F(t)}{t^{1-\gamma}}, \qquad \lambda \ge 0.$$

Without loss of generality, we assume that, for all $\lambda \ge m$, $\bar{F}(\lambda) = m$; otherwise, $\bar{F}$ can be replaced by the function $\bar{F} \wedge m$.

2.4.2 Coherence function

Suppose now that the spectral representation of S_* is $S_* = \sum_{k=1}^{r} \mu_k(\psi_k \otimes \psi_k)$, where $r = \operatorname{rank}(S_*) \ge 1$, μ_k are non-zero eigenvalues of S_* (possibly repeated) and ψ_k are

the corresponding orthonormal eigenfuctions. Let L be the range of S_* and P_L the orthogonal projection to L. The following function characterizes the relation between kernels S_* and W,

$$k \mapsto \sum_{j=1}^{k} \|P_L \phi_j\|^2$$

Ideally, we would like to express our bounds in terms of this function; nevertheless, due to some technicalities in the proof, we rely in a surrogate function φ such that $k \mapsto \frac{\varphi(k)}{\bar{F}(\lambda_k)}$ is nonincreasing and

$$\sum_{j=1}^{k} \|P_L \phi_j\|^2 \leq \varphi(k), \qquad k = 1, \ldots, m$$

It will be convenient to set $\varphi(k) = \varphi(m)$ for all $k \geq m$. We will denote by $\Psi = \Psi_{S_*,W}$ the class of all the functions satisfying these properties.

The following *coherence function* will be crucial in our analysis:

$$\bar{\varphi}(k) := \bar{\varphi}(S_*, k) := \max_{l \leq k} \bar{F}(\lambda_l) \max_{j \geq l} \frac{1}{\bar{F}(\lambda_j)} \sum_{j=1}^{k} \|P_L \phi_j\|^2,$$

$$k = 1, \ldots, m, \quad \bar{\varphi}(0) = 0.$$

It is straightforward to check that $\bar{\varphi} \in \Psi$ and, for all $\varphi \in \Psi$, $\bar{\varphi}(k) \leq \varphi(k), k = 0, \ldots, m$. Thus, $\bar{\varphi}$ is the smallest function $\varphi \in \Psi$. Also, $\bar{\varphi}(m) = r$ since $\sum_{j=1}^{m} \|P_L \phi_j\|^2 = \|P_L\|_F^2 = r$. Moreover, since $\frac{\bar{\varphi}(k)}{\bar{F}(\lambda_k)}$ is nonincreasing, we have

$$\bar{\varphi}(k) \geq \frac{r \bar{F}(\lambda_k)}{m}, \quad k = 0, \ldots, m.$$

The coherence function $\bar{\varphi}$ has some connection to the coherence constant used in noiseless low rank matrix completion problems. To be specific, when ν is the coherence of matrix S_* with respect to the basis $\{\phi_1, \ldots, \phi_m\}$, the following bound holds

$$\sum_{j=1}^{k} \|P_L \phi_j\|^2 \leq \frac{\nu r \bar{F}(\lambda_k)}{m}, k = 1, \ldots, m. \tag{16}$$

and thus

$$\bar{\varphi}(k) \leq \frac{\nu r \bar{F}(\lambda_k)}{m}, k = 1, \ldots, m.$$

which implies that condition (16) can be viewed as a weak version of low coherence.

2.4.3 Spectral characteristics on Erdős-Rényi graphs

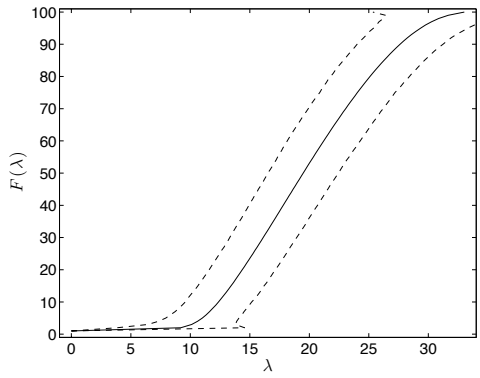
We illustrate the spectral characteristics of W and S_* on a problem of estimating smooth kernels over random graphs. For this purpose, we draw a random graph from the famous *Erdős-Rényi model* introduced independently by Edgar Gilbert [27], and by Paul Erdős and Alfréd Rényi [21]. In this model, we construct a random graph by including each possible edge at random with probability $p \in [0, 1]$ independently from every other edge. Equivalently, we consider the Erdős-Rényi model as a distribution $\mathcal{G}_{m,p}$ over graphs on m vertices, where the probability of a graph $G = ([m], E)$ is equal to $p^{|E|}(1-p)^{\binom{m}{2}-|E|}$. For a statistical analysis of the spectrum of random graphs, see [18, 23, 22, 5].

We estimate the spectral function $F(\lambda; \Delta)$ and its majorant $\bar{F}(\lambda; \Delta)$ where Δ is the Laplacian of a random graph $G = ([m], E)$. In figure 4, we show the expected value and confidence intervals for functions F and $\bar{F}$ in the case where G is sampled from $\mathcal{G}_{m,p}$, for $m = 100$ and different values of p. Let us remember that our goal is to find bounds in term of the spectral function F, but that due to technical reasons, we rely on the surrogate function $\bar{F}$. This implies that our bounds will be tighter when $\bar{F}$ is closer to F. In figure 5, we show F and $\bar{F}$ in the same plot for a better comparison. The smaller p is, the better $\bar{F}$ approximates F. The reason of this behavior is the spectral gap between the zero eigenvalue λ_1, and the first positive eigenvalue λ_{k_0}.

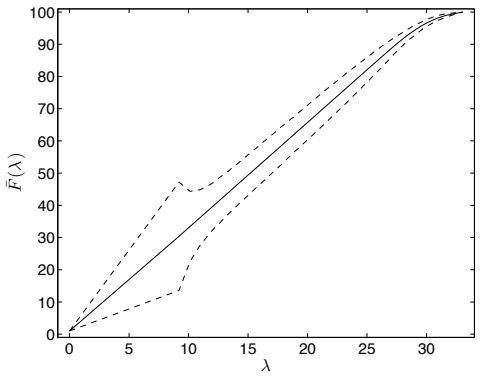
Let $G = ([m], E)$ be an Erdős-Rényi graph with Laplacian Δ with spectral representation $\Delta = \sum_{k=1}^{m} \lambda_k(\phi_k \otimes \phi_k)$. We construct a smooth and low-rank target kernel S_* over G by picking the parameters r_B and r_T in the following way,

$$S_* = S_*(r_B, r_T) := \sum_{k=1}^{r_B} (\phi_k \otimes \phi_k) + \sum_{k=m-r_T+1}^{m} \lambda_k^{-1} (\phi_k \otimes \phi_k)$$

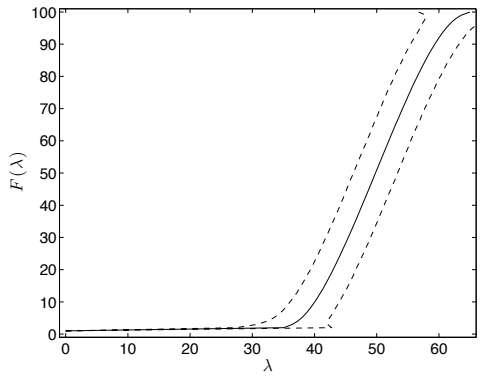
Note that, if we chose $r_B < k_0$ and $r_T \leq m - k_0 + 1$, then rank of S_* is equal to $\text{rank}(S_*) = r_B + r_T$, and the energy of S_* is equal to $\|\Delta^{1/2} S_*\|_F^2 = r_T$,



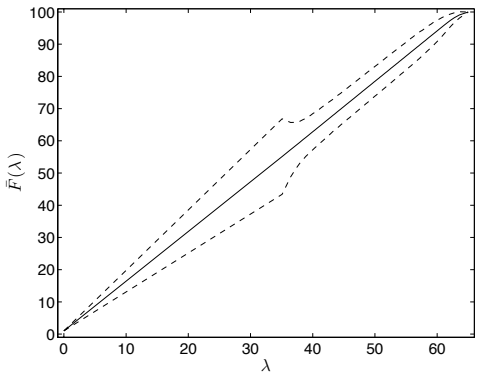
(a) $F(\lambda)$ for $p = 0.2$



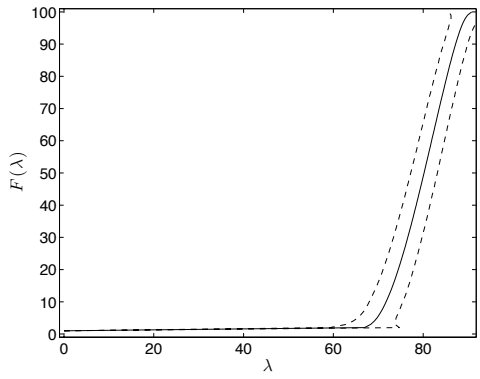
(b) $\bar{F}(\lambda)$ for $p = 0.2$



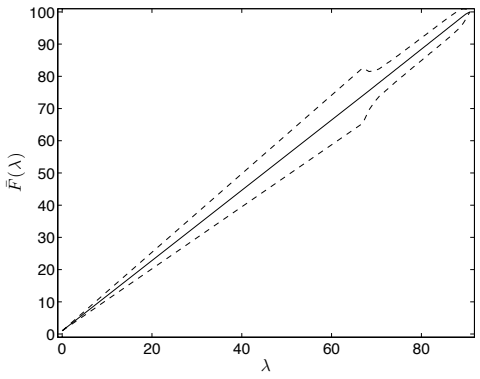
(c) $F(\lambda)$ for $p = 0.5$



(d) $\bar{F}(\lambda)$ for $p = 0.5$

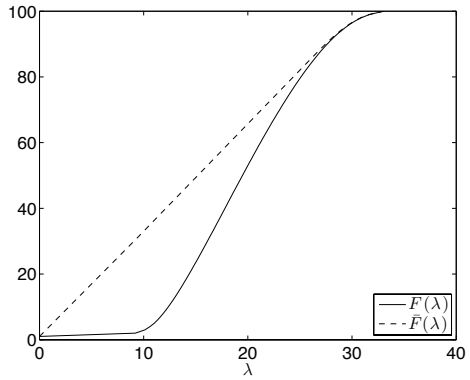


(e) $F(\lambda)$ for $p = 0.8$

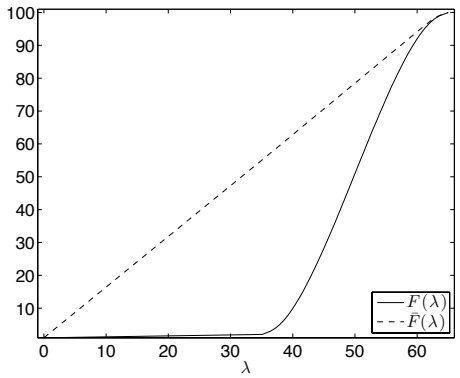


(f) $\bar{F}(\lambda)$ for $p = 0.8$

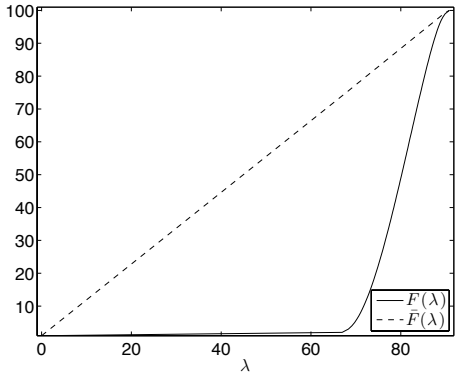
Figure 4: Mean value of the spectral function F and mean value of the majorant $\bar{F}$ for Erdős-Rényi graphs on 100 vertices and $p = 0.2$, 0.5 and 0.8



(a) $F(\lambda)$ and $\bar{F}(\lambda)$ for $p = 0.2$



(b) $F(\lambda)$ and $\bar{F}(\lambda)$ for $p = 0.5$



(c) $F(\lambda)$ and $\bar{F}(\lambda)$ for $p = 0.8$

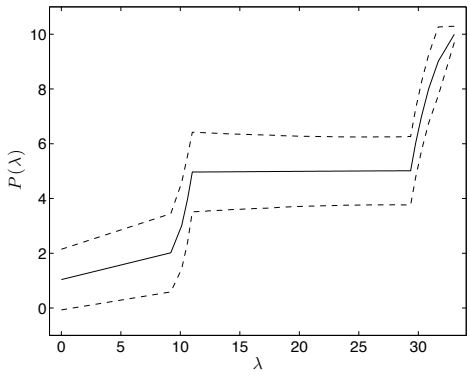
Figure 5: Comparison of spectral function F and its mayorant $\bar{F}$ for Erdős-Rényi graphs

In our estimation problem, we are interested in studying the following three coherent type functions,

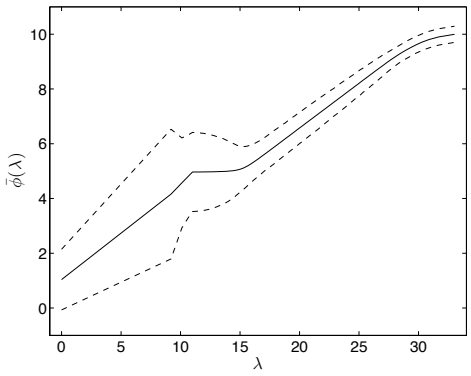
$$P(\lambda) = \sum_{\{k:\lambda_k<\lambda\}} \|P_L \phi_k\|^2, \qquad \bar{\phi}(\lambda_k) = \bar{\varphi}(k), \qquad \nu \frac{r}{m} \bar{F}(\lambda_k).$$

Remember that the majorant of $P(\lambda_k)$ is the coherent function $\bar{\phi}(\lambda)$ which is in turn bounded by $\nu \frac{r}{m} \bar{F}(\lambda_k)$. As explained in section 2.5, the proposed double penalty estimator performs better than the usual nuclear norm minimization estimator for matrices S_* with a large gap between $\bar{\phi}(\lambda_k)$ and $\nu \frac{r}{m} \bar{F}(\lambda_k)$. Moreover, the upper bound in theorem 2.1 is tighter when the gap between $P(\lambda_k)$ and $\bar{\phi}(\lambda_k)$ is small.

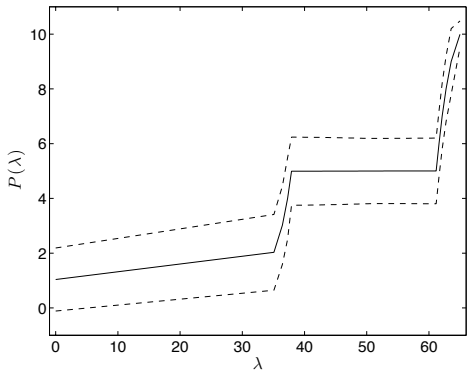
In figure 6, we show the expected value and confidence intervals for $P(\lambda)$ and $\bar{\phi}(\lambda)$ in the case where G is sampled from $\mathcal{G}_{m,p}$, for $m = 100$ and different values of p. For



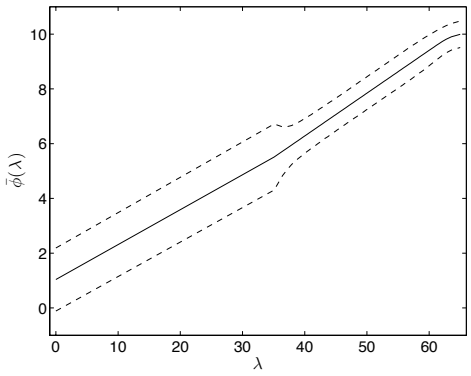
(a) $P(\lambda)$ for $p = 0.2$



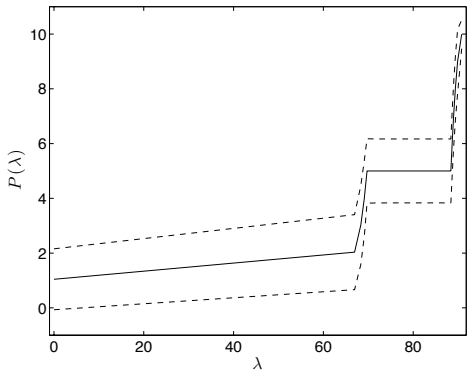
(b) $\bar{\phi}(\lambda)$ for $p = 0.2$



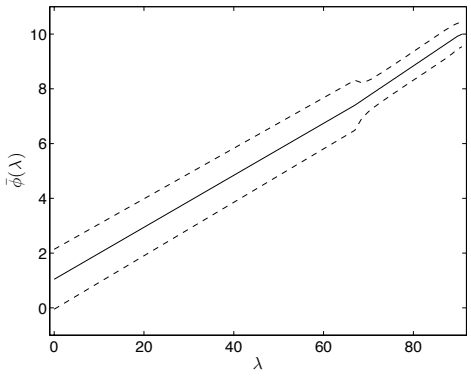
(c) $P(\lambda)$ for $p = 0.5$



(d) $\bar{\phi}(\lambda)$ for $p = 0.5$

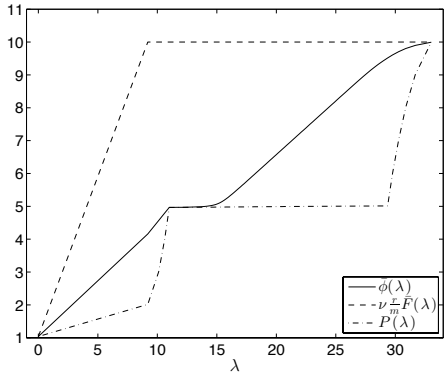


(e) $P(\lambda)$ for $p = 0.8$

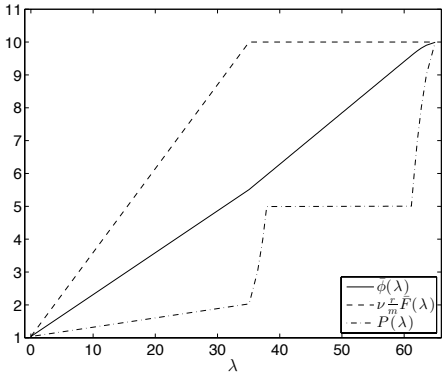


(f) $\bar{\phi}(\lambda)$ for $p = 0.8$

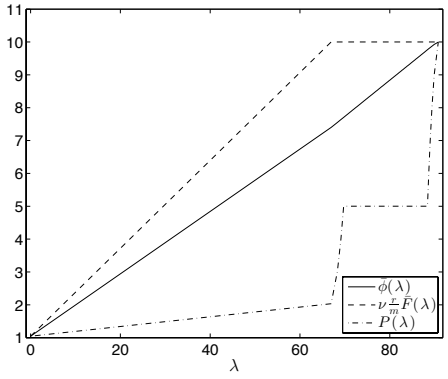
Figure 6: Mean value of the projection P and mean value of the coherence function $\bar{\phi}$ for Erdős-Rényi graphs on 100 vertices and $p = 0.2$, 0.5 and 0.8



(a) $P(\lambda)$ and $\bar{\phi}$ for $p = 0.2$



(b) $P(\lambda)$ and $\bar{\phi}$ for $p = 0.5$



(c) $P(\lambda)$ and $\bar{\phi}$ for $p = 0.8$

Figure 7: Comparison of projection P and the coherence function for Erdős-Rényi graphs

the construction of matrix S_*, we choose $r_B = r_T = 10$. In figure 7, we show the three coherence type functions in the same plot for a better comparison.

2.5 Analysis of the estimator

Given $t > 0$, we define $t_{n,m} := t + \log(2m \log_F(16 n^{\zeta} m^{(3/2)\zeta}))$. In what follows, we assume that $n \geq m t_{n,m}$. When $t \asymp \log m$, which is a typical choice of t, this assumption means that n should be larger than m times a log factor. We set the regularization parameter ε in (14) as:

$$\varepsilon := 4\sqrt{\frac{t + \log(2m)}{nm}}.$$

Theorem 2.1. There exist constants C, C_1 depending only on c such that, for all

$s \in \{k_0 + 1, \ldots, m + 1\}$ and all $\varepsilon_F \in [\lambda_s^{-1}, \lambda_{s-1}^{-1}]$, with probability at least $1 - e^{-t}$,

$$\begin{aligned}\frac{1}{m^2}\|\hat{S} - S_*\|_F^2 \leq C\frac{\bar{\varphi}(S_*; s)mt_{n,m}}{n} + \frac{\varepsilon_F}{m^2}\|W^{1/2}S_*\|_F^2 \\ + C_1 \max_{j=1,\ldots,m} \|P_L e_j\|^2 \Big(\frac{mt_{n,m}}{n}\Big)^2.\end{aligned} \tag{17}$$

Note that $\max_{v \in V} \|P_L e_v\|^2 \leq 1$. Thus, the last term in the righthand side of bound (17) is smaller than the first term, provided that

$$\bar{\varphi}(S_*; s) \geq \frac{mt_{n,m}}{n}$$

Moreover, this term is much smaller under a low coherence condition

$$\max_{v \in V} \|P_L e_v\|^2 \leq \frac{\nu r}{m}$$

for some $\nu \geq 1$. In this case,

$$\max_{v \in V} \|P_L e_v\|^2 \Big(\frac{mt_{n,m}}{n}\Big)^2 \leq \frac{\nu r m t_{n,m}^2}{n^2} \leq \frac{\nu r t_{n,m}}{n}.$$

Note also that Theorem 2.1 holds in the case when $\varepsilon_F = 0$. In this case, $s = m$ and $\bar{\varphi}(S_*, m) = r$, so the bound of Theorem 2.1 becomes

$$\frac{1}{m^2}\|\hat{S} - S_*\|_F^2 \leq C\frac{rmt_{n,m}}{n}, \tag{18}$$

which also follows from corollary 2 in [40].

Under condition (16), the following corollary of Theorem 2.1 holds.

Corollary 2.2. Suppose that condition (16) holds. Then, there exists a constant $C > 0$ depending only on ζ such that, for all $s \in \{k_0 + 1, \ldots, m + 1\}$ and all $\varepsilon_F \in (\lambda_s^{-1}, \lambda_{s-1}^{-1}]$, with probability at least $1 - e^{-t}$,

$$\begin{aligned}\frac{1}{m^2}\|\hat{S} - S_*\|_F^2 \leq C\frac{\nu r \bar{F}(\lambda_s)t_{n,m}}{n} + \frac{\varepsilon_F}{m^2}\|W^{1/2}S_*\|_F^2 \\ + C_1 \max_{v \in V} \|P_L e_v\|^2 \Big(\frac{mt_{n,m}}{n}\Big)^2.\end{aligned}$$

Note that, if $\lambda_k \asymp k^{2\beta}$ for some $\beta > 1/2$, then it is easy to see that one can choose $\bar{F}(\lambda) \asymp \lambda^{1/2\beta}$ and, with this choice, $\bar{F}(\lambda_s) \asymp s$. Thus, the value of s that minimizes the bound of Corollary 2.2 is

$$s \asymp \left(\frac{n}{\nu r m^2 t_{n,m}}\right)^{1/(2\beta+1)} \|W^{1/2} S_*\|_F^{2/(2\beta+1)},$$

which, under a low coherence assumption $\max_{v \in V} \|P_L e_v\|^2 \leq \frac{\nu r}{m}$, yields the bound

$$\|\hat{S} - S_*\|_F^2 \leq C\left(\frac{\nu r t_{n,m}}{n}\right)^{2\beta/(2\beta+1)} \|W^{1/2} S_*\|_F^{2/(2\beta+1)}. \tag{19}$$

The advantage of (19) comparing with (18) (that holds for $\varepsilon_F = 0$ and does not rely on any smoothness assumption on the kernel S_*) is due to the fact that there is no factor m in the numerator in the right hand side of (19). Due to this fact, when m is large enough and ν is not too large, bound (19) becomes sharper than (18).

2.6 Proof of Main Theorem

Given that the estimator $\hat{S}$ arises as the solution of a convex optimization problem, we begin the analysis by studying the sub-differential of the penalized empirical risk L_n. To do so, we need a characterization of the sub-differential of the nuclear norm. Such characterization is based on the following orthogonal projectors in the space $\mathcal{S}_{\mathcal{V}}$ with the Hilbert Schmidt inner product:

$$\mathcal{P}_L(A) := A - P_{L^\perp} A P_{L^\perp}, \qquad A \in \mathcal{S}_{\mathcal{V}}$$

$$\mathcal{P}_L^\perp(A) = P_{L^\perp} A P_{L^\perp}, \qquad A \in \mathcal{S}_{\mathcal{V}},$$

where $L \subseteq \mathbb{R}^{\mathcal{V}}$ is a given linear subspace, $L^\perp$ is its orthogonal complement, and P_L denotes the orthogonal projection to subspace L. Using these projections, we introduce the following well known representation of sub-differential of the convex function $S \mapsto \|S\|_*$ (see [60]):

$$\partial \|S\|_* = \left\{ \operatorname{sign}(S) + \mathcal{P}_L^\perp(M) : M \in \mathcal{S}_{\mathcal{V}},\, \|M\| \leq 1,\, L = \operatorname{range}(S) \right\},$$

Using this representation, we are able to proof the following bound,

Lemma 2.3. The following inequality holds for the estimator $\hat{S}$,

$$\begin{aligned} &\frac{2}{m^2}\|\hat{S}-S_*\|_F^2+\varepsilon\|\mathcal{P}_L^{\perp}(\hat{S})\|_*+\frac{2\varepsilon_F}{m^2}\|W^{1/2}(\hat{S}-S_*)\|_F^2 \\ &\leq -\varepsilon\langle \mathrm{sign}(S_*),\hat{S}-S_*\rangle-\frac{2\varepsilon_F}{m^2}\langle W^{1/2}S_*,W^{1/2}(\hat{S}-S_*)\rangle+2\langle \Xi,\hat{S}-S_*\rangle \end{aligned} \tag{20}$$

where

$$\Xi:=\frac{1}{m^2}(\breve{S}-S_*)$$

Proof. An arbitrary matrix $\hat{A}\in\partial L_n(\hat{S})$ can be represented as follows:

$$\hat{A}=\frac{2}{m^2}(\hat{S}-\breve{S})+\varepsilon\hat{V}+\frac{2\varepsilon_F}{m^2}W\hat{S}, \tag{21}$$

where $\hat{V}\in\partial\|\hat{S}\|_1$. Since $\hat{S}$ is a minimizer of $L_n(S)$, there exists a matrix $\hat{A}\in\partial L_n(\hat{S})$ such that $-\hat{A}$ belongs to the normal cone of $\mathbb{S}$ at the point $\hat{S}$. This implies that $\langle\hat{A},\hat{S}-S_*\rangle\leq 0$ and, in view of (21),

$$\frac{2}{m^2}\langle\hat{S}-\breve{S},\hat{S}-S_*\rangle+\varepsilon\langle\hat{V},\hat{S}-S_*\rangle+\frac{2\varepsilon_F}{m^2}\langle W\hat{S},\hat{S}-S_*\rangle\geq 0$$

It follows by a simple algebra that

$$\begin{aligned} &\frac{2}{m^2}\|\hat{S}-S_*\|_F^2+\frac{2\varepsilon_F}{m^2}\|W^{1/2}(\hat{S}-S_*)\|_F^2+\varepsilon\langle\hat{V},\hat{S}-S_*\rangle \\ &\qquad\leq -\frac{2\varepsilon_F}{m^2}\langle S_*,W(\hat{S}-S_*)\rangle+2\langle\Xi,\hat{S}-S_*\rangle, \end{aligned} \tag{22}$$

On the other hand, let $V_*\in\partial\|S_*\|_*$. Therefore, the representation $V_*=\mathrm{sign}(S_*)+\mathcal{P}_L^{\perp}(M)$ holds, where M is a matrix with $\|M\|\leq 1$. It follows from the trace duality property that there exists an M with $\|M\|\leq 1$ such that

$$\langle\mathcal{P}_L^{\perp}(M),\hat{S}-S_*\rangle=\langle M,\mathcal{P}_L^{\perp}(\hat{S}-S_*)\rangle=\langle M,\mathcal{P}_L^{\perp}(\hat{S})\rangle=\|\mathcal{P}_L^{\perp}(\hat{S})\|_*$$

where in the first equality we used that $\mathcal{P}_L^{\perp}$ is a self-adjoint operator and in the second equality we used that S_* has range L. Using this equation and monotonicity of subdifferentials of convex functions, we get

$$\langle\mathrm{sign}(S_*),\hat{S}-S_*\rangle+\|\mathcal{P}_L^{\perp}(\hat{S})\|_*=\langle V_*,\hat{S}-S\rangle\leq\langle\hat{V},\hat{S}-S_*\rangle$$

Substituting this in (22), we get the result □

The rest of the proof consists on bounding each term in the right hand side of (20) in terms of an arbitrary function $\varphi \in \Psi_{S_*,W}$ with $\varphi(k) = r, k \geq m$. Then we get the main result (17) by substituting φ for $\bar{\varphi}$ which is the smallest function in $\Psi_{S_*,W}$. Throughout the proof, we assume that $s \in \{k_0, \ldots, m\}$ and $\varepsilon_F \in [\lambda_{s+1}^{-1}, \lambda_s^{-1}]$ (at the end of the proof, we replace $s+1 \mapsto s$).

2.6.1 Bounding the first term

First note that

$$\begin{aligned} \varepsilon|\langle \mathrm{sign}(S_*), \hat{S} - S_* \rangle| &\leq \varepsilon \|\mathrm{sign}(S_*)\|_F \|\hat{S} - S_*\|_F \\ &= \varepsilon\sqrt{r}\|\hat{S} - S_*\|_F \leq \frac{1}{2} r m^2 \varepsilon^2 + \frac{1}{2m^2}\|\hat{S} - S_*\|_F^2. \end{aligned} \tag{23}$$

We will also need a more subtle bound on $\langle \mathrm{sign}(S_*), \hat{S} - S_* \rangle$, expressed in terms of function φ. Note that, for all $k_0 \leq s \leq m$,

$$\begin{aligned} \langle \mathrm{sign}(S_*), \hat{S} - S_* \rangle &= \sum_{k=1}^{m} \langle \mathrm{sign}(S_*)\phi_k, (\hat{S} - S_*)\phi_k \rangle \\ = \sum_{k=1}^{s} \langle \mathrm{sign}(S_*)\phi_k, (\hat{S} - S_*)\phi_k \rangle &+ \sum_{k=s+1}^{m} \left\langle \frac{\mathrm{sign}(S_*)\phi_k}{\sqrt{\lambda_k}}, \sqrt{\lambda_k}(\hat{S} - S_*)\phi_k \right\rangle, \end{aligned}$$

which easily implies

$$\begin{aligned} &|\langle \mathrm{sign}(S_*), \hat{S} - S_* \rangle| \\ &\leq \left(\sum_{k=1}^{s} \|\mathrm{sign}(S_*)\phi_k\|^2 \right)^{1/2} \left(\sum_{k=1}^{s} \|(\hat{S} - S_*)\phi_k\|^2 \right)^{1/2} + \\ &\left(\sum_{k=s+1}^{m} \frac{\|\mathrm{sign}(S_*)\phi_k\|^2}{\lambda_k} \right)^{1/2} \left(\sum_{k=s+1}^{m} \lambda_k \|(\hat{S} - S_*)\phi_k\|^2 \right)^{1/2} \\ &\leq \left(\sum_{k=1}^{s} \|P_L \phi_k\|^2 \right)^{1/2} \|\hat{S} - S_*\|_F \\ &+ \left(\sum_{k=s+1}^{m} \frac{\|P_L \phi_k\|^2}{\lambda_k} \right)^{1/2} \|W^{1/2}(\hat{S} - S_*)\|_F. \end{aligned} \tag{24}$$

We will now use the following elementary lemma.

Lemma 2.4. Let c, γ be the constants involved in the conditions on the spectrum of W and in the definition of $\bar{F}$. For all $s \geq k_0 - 1$,

$$\sum_{k=s+1}^{m} \frac{\|P_L \phi_k\|^2}{\lambda_k} \leq c_\gamma \frac{\varphi(s+1)}{\lambda_{s+1}} \quad \text{and} \quad \sum_{k=s+1}^{m} \frac{1}{\lambda_k} \leq c_\gamma \frac{\bar{F}(\lambda_{s+1})}{\lambda_{s+1}},$$

where $c_\gamma := \frac{c}{\gamma} + 1$.

Proof. Denote $F_s := \sum_{k=1}^{s} \|P_L \phi_k\|^2, s = 1, \dots, m$. Then, using the properties of functions $\varphi \in \Psi$ and $\bar{F}$, and of the spectrum of W, we get

$$\begin{aligned}
\sum_{k=s+1}^{m} \frac{\|P_L \phi_k\|^2}{\lambda_k} &= \sum_{k=s+1}^{m-1} F_k \Big(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k+1}} \Big) + \frac{F_m}{\lambda_m} - \frac{F_s}{\lambda_{s+1}} \\
&\le \sum_{k=s+1}^{m-1} \varphi(k) \Big(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k+1}} \Big) + \frac{\varphi(m)}{\lambda_m} \\
&\le \frac{\varphi(s+1)}{\bar{F}(\lambda_{s+1})} \Big[\sum_{k=s+1}^{m-1} \frac{\bar{F}(\lambda_k)}{\lambda_k \lambda_{k+1}} (\lambda_{k+1} - \lambda_k) + \frac{\bar{F}(\lambda_m)}{\lambda_m} \Big] \\
&\le c \frac{\varphi(s+1)}{\bar{F}(\lambda_{s+1})} \sum_{k=s+1}^{m-1} \frac{\bar{F}(\lambda_{k+1})}{\lambda_{k+1}^2} (\lambda_{k+1} - \lambda_k) + \frac{\varphi(s+1)}{\bar{F}(\lambda_{s+1})} \frac{\bar{F}(\lambda_{s+1})}{\lambda_{s+1}} \\
&\le c \frac{\varphi(s+1)}{\bar{F}(\lambda_{s+1})} \int_{\lambda_{s+1}}^{\infty} \frac{\bar{F}(t)}{t^2} dt + \frac{\varphi(s+1)}{\lambda_{s+1}} \\
&\le \frac{c}{\gamma} \frac{\varphi(s+1)}{\bar{F}(\lambda_{s+1})} \frac{\bar{F}(\lambda_{s+1})}{\lambda_{s+1}} + \frac{\varphi(s+1)}{\lambda_{s+1}} = c_\gamma \frac{\varphi(s+1)}{\lambda_{s+1}}.
\end{aligned} \tag{25}$$

The proof of the second bound is similar (with some simplifications). □

It follows from from (24) and the bound of Lemma 2.4 that

$$\begin{aligned}
&|\langle \operatorname{sign}(S_*), \hat{S} - S_* \rangle| \\
&\le \sqrt{\varphi(s)} \|\hat{S} - S_*\|_F + \sqrt{c_\gamma \frac{\varphi(s+1)}{\lambda_{s+1}}} \|W^{1/2}(\hat{S} - S_*)\|_F
\end{aligned} \tag{26}$$

This implies the following bound:

$$\begin{aligned}
&\varepsilon |\langle \operatorname{sign}(S_*), \hat{S} - S_* \rangle| \\
&\le \varphi(s) m^2 \varepsilon^2 + \frac{1}{4m^2} \|\hat{S} - S_*\|_F^2 \\
&+ c_\gamma \frac{\varphi(s+1)}{\lambda_{s+1}} \frac{m^2 \varepsilon^2}{\varepsilon_F} + \frac{\varepsilon_F}{4m^2} \|W^{1/2}(\hat{S} - S_*)\|_F^2,
\end{aligned} \tag{27}$$

where we used twice an elementary inequality $ab \le a^2 + \frac{1}{4} b^2, a, b > 0$. Since, under the assumptions of the theorem, $\varepsilon_F \lambda_{s+1} \ge 1$, inequality (27) yields the following bound:

$$\begin{aligned}
\varepsilon |\langle \operatorname{sign}(S_*), \hat{S} - S_* \rangle| &\le (c_\gamma + 1) \varphi(s+1) m^2 \varepsilon^2 \\
&+ \frac{1}{4} \|\hat{S} - S_*\|_F^2 + \frac{\varepsilon_F}{4m^2} \|W^{1/2}(\hat{S} - S_*)\|_F^2.
\end{aligned} \tag{28}$$

2.6.2 Bounding the second term

To bound the second term in the right hand side of (20), note that

$$|\langle W^{1/2}S_*, W^{1/2}(\hat{S}-S_*)\rangle| \le \|W^{1/2}S_*\|_F \|W^{1/2}(\hat{S}-S_*)\|_F, \tag{29}$$

which implies

$$\begin{aligned} \varepsilon_F |\langle W^{1/2}S_*, W^{1/2}(\hat{S}-S_*)\rangle| &\le \varepsilon_F \|W^{1/2}S_*\|_F^2 + \frac{\varepsilon_F}{4}\|W^{1/2}(\hat{S}-S_*)\|_F^2 \\ &= \frac{\varepsilon_F}{m^2}\|W^{1/2}S_*\|_F^2 + \frac{\varepsilon_F}{4m^2}\|W^{1/2}(\hat{S}-S_*)\|_F^2. \end{aligned} \tag{30}$$

2.6.3 Bounding the third term

Finally, we bound $\langle \Xi, \hat{S}-S_*\rangle$:

$$\begin{aligned} |\langle \Xi, \hat{S}-S_*\rangle| &\le |\langle \Xi, \mathcal{P}_L(\hat{S}-S_*)\rangle| + |\langle \Xi, \mathcal{P}_L^{\perp}(\hat{S})\rangle| \\ &\le |\langle \mathcal{P}_L\Xi, \hat{S}-S_*\rangle| + \|\Xi\|\|\mathcal{P}_L^{\perp}(\hat{S})\|_* \end{aligned} \tag{31}$$

To bound $\|\Xi\|$, we use a version of noncommutative Bernstein inequality of Ahlswede and Winter [1]. Other versions of this kind of inequalities can be found in [58] and [38].

Lemma 2.5. Let Z be a bounded random symmetric matrix with $\mathbb{E}Z = 0$, $\sigma_Z^2 := \|\mathbb{E}Z^2\|$ and $\|Z\| \le M$ for some $M > 0$. Let $Z_1, \dots, Z_n$ be n i.i.d. copies of Z. Then for all $t > 0$, with probability at least $1 - e^t$

$$\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\| \le 2\left(\sigma_Z\sqrt{\frac{t+\log(2m)}{n}} \bigvee M\frac{t+\log(2m)}{n}\right)$$

It is applied to i.i.d. random matrices $Z_i := Y_i(e_U \otimes e_V) - \mathbb{E}(Y_i(e_U \otimes e_V)), i = 1, \dots, n$. Since $\|Z_i\| \le 2$ and, by a simple computation, $\sigma_{Z_i}^2 := \|\mathbb{E}Z_i^2\| \le 1/m$ (see Section 9.4 in [38]), Lemma 2.5 implies that with probability at least $1 - e^{-t}$

$$\|\Xi\| = \left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\| \le 2\left[\sqrt{\frac{t+\log(2m)}{nm}} \bigvee \frac{2(t+\log(2m))}{n}\right].$$

Under the assumption that

$$\varepsilon \ge 4\left[\sqrt{\frac{t+\log(2m)}{nm}} \bigvee \frac{2(t+\log(2m))}{n}\right],$$

this yields $\|\Xi\| \leq \varepsilon/2$ and

$$|\langle \Xi, \hat{S} - S_* \rangle| \leq |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| + \frac{\varepsilon}{2} \|\mathcal{P}_L^{\perp}(\hat{S})\|_*. \tag{32}$$

For simplicity, it is assumed that $n \geq 2m(t + \log(2m))$. In this case, one can take $\varepsilon = 4\sqrt{\frac{t+\log(2m)}{nm}}$, as it has been done in the statement of the theorem.

We have to bound $|\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle|$ and we start with the following simple bound:

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| &\leq \|\mathcal{P}_L \Xi\|_F \|\hat{S} - S_*\|_F \\ &\leq \sqrt{2r} \|\Xi\| \|\hat{S} - S_*\|_F \\ &\leq \frac{1}{2} \varepsilon \sqrt{2r} \|\hat{S} - S_*\|_F \\ &\leq \frac{1}{2} m^2 \varepsilon^2 r + \frac{1}{4m^2} \|\hat{S} - S_*\|_F^2, \end{aligned} \tag{33}$$

where we use the fact that $\operatorname{rank}(\mathcal{P}_L \Xi) \leq 2r$. Substituting (23), (30), (32) and (33) in (20), we easily get that

$$\|\hat{S} - S_*\|_F^2 \leq \frac{3}{2} r \varepsilon^2 m^2 + 2 \frac{\varepsilon_F}{m^2} \|W^{1/2} S_*\|_F^2. \tag{34}$$

For $\bar{\varepsilon} = 0$, this bound follows from the results of Koltchinskii, Lounici and Tsybakov (2011). However, we need a more subtle bound expressed in terms of function φ, which is akin to bound (28). To this end, we will use the following lemma.

Lemma 2.6. For $\delta > 0$, let $k(\delta) := F(\delta^{-2})$ (that is, $k(\delta)$ is the largest value of $k \leq m$ such that $\lambda_k^{-1} \geq \delta^2$). For all $t > 0$, with probability at least $1 - e^{-t}$,

$$\begin{aligned} \sup_{\|M\|_F \leq \delta, \|W^{1/2} M\|_F \leq 1} |\langle \mathcal{P}_L \Xi, M \rangle| &\leq 2\sqrt{4(c_\gamma + 1)} \sqrt{\frac{t}{nm}} \delta \sqrt{\varphi(k(\delta) + 1)} \\ &\quad + 2\sqrt{2} \delta \max_{v \in V} \|P_L e_v\| \frac{t}{n}, \end{aligned}$$

provided that $k(\delta) < m$, and

$$|\langle \mathcal{P}_L \Xi, M \rangle| \leq 4\sqrt{2} \delta \sqrt{\frac{rt}{nm}} + 2\sqrt{2} \delta \max_{v \in V} \|P_L e_v\| \frac{t}{n},$$

provided that $k(\delta) \geq m$.

Proof. The proof is somewhat akin to the derivation of the bounds on Rademacher processes in terms of Mendelson's complexities used in learning theory (see Proposition 3.3 in [38]). Note that, for all symmetric $m \times m$ matrices M,

$$\langle \mathcal{P}_L \Xi, M \rangle = \sum_{k,j=1}^{m} \langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j \rangle \langle M, \phi_k \otimes \phi_j \rangle.$$

Suppose that

$$\|M\|_F^2 = \sum_{k,j=1}^{m} |\langle M, \phi_k \otimes \phi_j \rangle|^2 \leq \delta^2$$

and

$$\|W^{1/2} M\|_F^2 = \sum_{k,j=1}^{m} \lambda_k |\langle M, \phi_k \otimes \phi_j \rangle|^2 \leq 1.$$

Then, it easily follows that

$$\sum_{k,j=1}^{m} \frac{|\langle M, \phi_k \otimes \phi_j \rangle|^2}{\lambda_k^{-1} \wedge \delta^2} \leq 2,$$

which implies

$$\begin{aligned} & |\langle \mathcal{P}_L \Xi, M \rangle|^2 \\ & \leq \sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j \rangle|^2 \sum_{k,j=1}^{m} \frac{|\langle M, \phi_k \otimes \phi_j \rangle|^2}{\lambda_k^{-1} \wedge \delta^2} \\ & \leq 2 \sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j \rangle|^2 \end{aligned} \tag{35}$$

Define now the following inner product:

$$\langle M_1, M_F \rangle_w := \sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2) \langle M_1, \phi_k \otimes \phi_j \rangle \langle M_F, \phi_k \otimes \phi_j \rangle$$

and let $\| \cdot \|_w$ be the corresponding norm. We will provide an upper bound on

$$\|\mathcal{P}_L \Xi\|_w = \left(\sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j \rangle|^2 \right)^{1/2}.$$

To this end, we use a standard Bernstein type inequality for random variables in a Hilbert space. It is given in the following lemma (which follows, for instance, from Theorem 3.3.4(b) in [62]).

Lemma 2.7. Let ξ be a bounded random variable with values in a Hilbert space $\mathcal{H}$. Suppose that $\mathbb{E}\xi = 0$, $\mathbb{E}\|\xi\|_{\mathcal{H}}^2 = \sigma^2$ and $\|\xi\|_{\mathcal{H}} \le M$. Let $\xi_1, \dots, \xi_n$ be n i.i.d. copies of ξ_i. Then for all $t > 0$, with probability at least $1 - e^t$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\|_{\mathcal{H}} \le 2 \left[\sigma \sqrt{\frac{t}{n}} \bigvee M \frac{t}{n} \right]$$

Applying Lemma 2.7 to the random variable $\xi = Y\mathcal{P}_L(e_U \otimes e_V) - \mathbb{E}Y\mathcal{P}_L(e_U \otimes e_V)$, we get that for all $t > 0$, with probability at least $1 - e^{-t}$,

$$\begin{aligned} \|\mathcal{P}_L \Xi\|_w &= \left\| \frac{1}{n} \sum_{j=1}^{n} Y_j \mathcal{P}_L(e_{U_j} \otimes e_{V_j}) - \mathbb{E}Y\mathcal{P}_L(e_U \otimes e_V) \right\|_w \\ &\le 2 \left[\mathbb{E}^{1/2} \|Y\mathcal{P}_L(e_U \otimes e_V)\|_w^2 \sqrt{\frac{t}{n}} + \left\| \|Y\mathcal{P}_L(e_U \otimes e_V)\|_w \right\|_{L_\infty} \frac{t}{n} \right]. \end{aligned} \tag{36}$$

Using the fact that $Y \in \{-1, 1\}$, we get

$$\begin{aligned} \mathbb{E}\|Y\mathcal{P}_L(e_U \otimes e_V)\|_w^2 &= \mathbb{E}\|\mathcal{P}_L(e_U \otimes e_V)\|_w^2 \\ &= \mathbb{E} \sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L(e_U \otimes e_V), \phi_k \otimes \phi_j \rangle|^2 \\ &= \sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2) \mathbb{E} |\langle e_U \otimes e_V, \mathcal{P}_L(\phi_k \otimes \phi_j) \rangle|^2 \\ &= \sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2) m^{-2} \sum_{u,v \in V} |\langle e_u \otimes e_v, \mathcal{P}_L(\phi_k \otimes \phi_j) \rangle|^2 \\ &\le m^{-2} \sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2) \|\mathcal{P}_L(\phi_k \otimes \phi_j)\|_F^2 \\ &\le 2m^{-2} \sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2) (\|P_L \phi_k\|^2 + \|P_L \phi_j\|^2) \\ &= 2m^{-1} \sum_{k=1}^{m} (\lambda_k^{-1} \wedge \delta^2) \|P_L \phi_k\|^2 + 2m^{-2} \sum_{k=1}^{m} (\lambda_k^{-1} \wedge \delta^2) \sum_{j=1}^{m} \|P_L \phi_j\|^2 \\ &= 2m^{-1} \sum_{k=1}^{m} (\lambda_k^{-1} \wedge \delta^2) \|P_L \phi_k\|^2 + 2m^{-2} \sum_{k=1}^{m} (\lambda_k^{-1} \wedge \delta^2) \|P_L\|_F^2 \\ &= 2m^{-1} \sum_{k=1}^{m} (\lambda_k^{-1} \wedge \delta^2) \|P_L \phi_k\|^2 + 2m^{-2} r \sum_{k=1}^{m} (\lambda_k^{-1} \wedge \delta^2). \end{aligned} \tag{37}$$

To bound $\mathbb{E}\|Y\mathcal{P}_L(e_U \otimes e_V)\|_w^2$ further, note that

$$\sum_{k=1}^{m} (\lambda_k^{-1} \wedge \delta^2) \|P_L \phi_k\|^2 \le \delta^2 \sum_{k \le k(\delta)} \|P_L \phi_k\|^2 + \sum_{k > k(\delta)} \lambda_k^{-1} \|P_L \phi_k\|^2. \tag{38}$$

Assuming that $1 \le k(\delta) \le m-1$, using the first bound of Lemma 2.4, the fact that $\lambda_{k(\delta)+1}^{-1} < \delta^2$ and the monotonicity of function φ, we get from (38) that

$$\begin{aligned}\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\|P_L\phi_k\|^2 &\le \delta^2\varphi(k(\delta)) + c_\gamma\frac{\varphi(k(\delta)+1)}{\lambda_{k(\delta)+1}} \\ &\le \delta^2\varphi(k(\delta)) + c_\gamma\delta^2\varphi(k(\delta)+1) \le (c_\gamma+1)\delta^2\varphi(k(\delta)+1).\end{aligned} \tag{39}$$

It is easy to check that (39) holds also for $k(\delta)=0$ and $k(\delta)=m$ (in the last case, $\varphi(k(\delta)+1)=r$). We also have

$$\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2) \le \sum_{k\le k(\delta)}\delta^2 + \sum_{k>k(\delta)}\lambda_k^{-1},$$

which, in view of the second bound of Lemma 2.4 and the properties of function φ, implies

$$\begin{aligned}\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2) &\le \delta^2 k(\delta) + c_\gamma\frac{\bar F(\lambda_{k(\delta)+1})}{\lambda_{k(\delta)+1}} \\ &\le (c_\gamma+1)\delta^2\bar F(\lambda_{k(\delta)+1}) \le (c_\gamma+1)\frac{m}{r}\delta^2\varphi(k(\delta)+1).\end{aligned} \tag{40}$$

Using bounds (37), (39) and (40), we get, under the condition that $k(\delta)<m$,

$$\begin{aligned}&\mathbb{E}\|Y\mathcal{P}_L(e_U\otimes e_V)\|_w^2 \\ &\le 2m^{-1}(c_\gamma+1)\delta^2\varphi(k(\delta)+1) + 2m^{-2}r(c_\gamma+1)\frac{m}{r}\delta^2\varphi(k(\delta)+1) \\ &\le 4(c_\gamma+1)m^{-1}\delta^2\varphi(k(\delta)+1).\end{aligned} \tag{41}$$

In the case when $k(\delta)\ge m$, it is easy to show that

$$\mathbb{E}\|Y\mathcal{P}_L(e_U\otimes e_V)\|_w^2 \le 4m^{-1}\delta^2 r. \tag{42}$$

We can also bound $\Big\|\|Y\mathcal{P}_L(e_U\otimes e_V)\|_w\Big\|_{L_\infty}^2$ as follows:

$$\begin{aligned}\Big\|\|Y\mathcal{P}_L(e_U\otimes e_V)\|_w\Big\|_{L_\infty}^2 &= \Big\|\|\mathcal{P}_L(e_U\otimes e_V)\|_w\Big\|_{L_\infty}^2 \\ &= \Big\|\sum_{k,j=1}^{m}(\lambda_k^{-1}\wedge\delta^2)|\langle\mathcal{P}_L(e_U\otimes e_V),\phi_k\otimes\phi_j\rangle|^2\Big\|_{L_\infty} \\ &\le \max_{1\le k\le m}(\lambda_k^{-1}\wedge\delta^2)\max_{u,v\in V}\sum_{k,j=1}^{m}|\langle\mathcal{P}_L(e_u\otimes e_v),\phi_k\otimes\phi_j\rangle|^2 \\ &\le \max_{1\le k\le m}(\lambda_k^{-1}\wedge\delta^2)\max_{u,v\in V}\|\mathcal{P}_L(e_u\otimes e_v)\|_F^2 \\ &\le \delta^2\max_{u,v\in V}\|\mathcal{P}_L(e_u\otimes e_v)\|_F^2 \le 2\delta^2\max_{v\in V}\|P_Le_v\|^2.\end{aligned} \tag{43}$$

If $k(\delta) < m$, it follows from (35), (36), (41) and (43) that with probability at least $1 - e^{-t}$, for all symmetric matrices M with $\|M\|_F \le \delta$ and $\|W^{1/2}M\|_F \le 1$,

$$|\langle \mathcal{P}_L \Xi, M\rangle| \le 2\sqrt{4(c_\gamma+1)}\sqrt{\frac{t}{nm}}\delta\sqrt{\varphi(k(\delta)+1)} + 2\sqrt{2}\delta \max_{v\in V}\|P_L e_v\|\frac{t}{n}.$$

Alternatively, if $k(\delta) \ge m$, we use (42) to get

$$|\langle \mathcal{P}_L \Xi, M\rangle| \le 4\delta\sqrt{\frac{rt}{nm}} + 2\sqrt{2}\delta \max_{v\in V}\|P_L e_v\|\frac{t}{n}.$$

This completes the proof of Lemma 2.6. □

It follows from Lemma 2.6 that, for all $\delta > 0$, the following bound holds with probability at least $1 - e^{-t}$

$$\begin{aligned} &\sup_{\|M\|_F\le\delta, \|W^{1/2}M\|_F\le 1} |\langle \mathcal{P}_L \Xi, M\rangle| \\ &\le 4\sqrt{c_\gamma+1}\sqrt{\frac{t}{nm}}\delta\sqrt{\varphi(k(\delta)+1)} + 2\sqrt{2}\delta \max_{v\in V}\|P_L e_v\|\frac{t}{n} \end{aligned} \tag{44}$$

(recall that $\varphi(k) = r$ for $k \ge m$, so, the second bound of the lemma can be included in the first bound). Moreover, the bound can be easily made uniform in $\delta \in [\delta_-, \delta_+]$ for arbitrary $\delta_- < \delta_+$. To this end, take $\delta_j := \delta_+ 2^{-j}, j = 0, 1, \dots [\log_F(\delta_+/\delta_-)] + 1$ and use (44) for each $\delta = \delta_j$ with $\bar{t} := t + \log([\log_F(\delta_+/\delta_-)] + 2)$ instead of t. An application of the union bound and monotonicity of the left hand side and the right hand side of (44) with respect to δ then implies that with probability at least $1 - e^{-t}$ for all $\delta \in [\delta_-, \delta_+]$

$$\begin{aligned} &\sup_{\|M\|_F\le\delta, \|W^{1/2}M\|_F\le 1} |\langle \mathcal{P}_L \Xi, M\rangle| \\ &\le C\sqrt{\frac{\bar{t}}{nm}}\delta\sqrt{\varphi(k(\delta)+1)} + 4\sqrt{2}\delta \max_{v\in V}\|P_L e_v\|\frac{\bar{t}}{n}. \end{aligned} \tag{45}$$

where $C > 0$ is a constant depending only on c. Indeed, by the union bound, (44) holds with probability at least

$$1 - ([\log_F(\delta_+/\delta_-)] + 2)e^{-\bar{t}} = 1 - e^{-t}$$

for all $\delta = \delta_j, j = 0, \dots, [\log_F(\delta_+/\delta_-)] + 1$.

Therefore, for all $j = 0, \dots, [\log_F(\delta_+/\delta_-)] + 1$ and all $\delta \in (\delta_{j+1}, \delta_j]$

$$\begin{aligned} &\sup_{\|M\|_F \le \delta, \|W^{1/2}M\|_F \le 1} |\langle \mathcal{P}_L \Xi, M \rangle| \\ &\le 4\sqrt{c_\gamma + 1}\sqrt{\frac{\bar{t}}{nm}}\delta_j \sqrt{\varphi(k(\delta_j)+1)} + 2\sqrt{2}\delta_j \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n} \end{aligned} \tag{46}$$

(by monotonicity of the left hand side). Note that $k(\delta_j) \le k(\delta) \le k(\delta_{j+1})$. We can now use the fact that $\frac{\varphi(k)}{\lambda_k} = \frac{\varphi(k)}{\bar{F}(\lambda_k)} \frac{\bar{F}(\lambda_k)}{\lambda_k}$ is a nonincreasing function and the condition $\lambda_{k+1}/\lambda_k \le c$ to show that

$$\begin{aligned} &\sqrt{\frac{\bar{t}}{nm}}\delta_j \sqrt{\varphi(k(\delta_j)+1)} + \le 2\sqrt{\frac{\bar{t}}{nm}}\delta_{j+1}\sqrt{\varphi(k(\delta_{j+1})+1)} \\ &\le 2\sqrt{\frac{\bar{t}}{nm}}\sqrt{\frac{\varphi(k(\delta_{j+1})+1)}{\lambda_{k(\delta_{j+1})}}} \le 2\sqrt{c}\sqrt{\frac{\bar{t}}{nm}}\sqrt{\frac{\varphi(k(\delta_{j+1})+1)}{\lambda_{k(\delta_{j+1})+1}}} \\ &\le 2\sqrt{c}\sqrt{\frac{\bar{t}}{nm}}\sqrt{\frac{\varphi(k(\delta)+1)}{\lambda_{k(\delta)+1}}} \le 2\sqrt{c}\sqrt{\frac{\bar{t}}{nm}}\delta\sqrt{\varphi(k(\delta)+1)}. \end{aligned}$$

This and bound (46) imply that

$$\begin{aligned} &\sup_{\|M\|_F \le \delta, \|W^{1/2}M\|_F \le 1} |\langle \mathcal{P}_L \Xi, M \rangle| \\ &\le 8\sqrt{c(c_\gamma + 1)}\sqrt{\frac{\bar{t}}{nm}}\delta \sqrt{\varphi(k(\delta)+1)} + 4\sqrt{2}\delta \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n}, \end{aligned} \tag{47}$$

which proves bound (45).

Set δ as

$$\delta := \frac{\|\hat{S} - S_*\|_F}{\|W^{1/2}(\hat{S} - S_*)\|_F}$$

and assume for now that $\delta \in [\delta_-, \delta_+]$. For a particular choice of $M := \frac{\hat{S} - S_*}{\|W^{1/2}(\hat{S} - S_*)\|_F}$, we get from (45) that

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| &\le C\sqrt{\frac{\bar{t}}{nm}}\|\hat{S} - S_*\|_F \sqrt{\varphi(k(\delta)+1)} \\ &+ 4\sqrt{2} \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n} \|\hat{S} - S_*\|_F. \end{aligned} \tag{48}$$

Suppose now that $\delta^2 \geq \varepsilon_F$. Since, under assumptions of the theorem, $\varepsilon_F \in (\lambda_{s+1}^{-1}, \lambda_s^{-1}]$, this implies that $k(\delta) \leq k(\sqrt{\varepsilon_F}) = s$ and

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| &\leq C \sqrt{\frac{\bar{t}}{nm}} \|\hat{S} - S_*\|_F \sqrt{\varphi(s+1)} \\ &+ 4\sqrt{2} \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n} \|\hat{S} - S_*\|_F \\ &\leq 2C^2 \frac{\varphi(s+1) m \bar{t}}{n} + 64 \max_{v \in V} \|P_L e_v\|^2 \Big(\frac{m\bar{t}}{n} \Big)^2 + \frac{1}{4m^2} \|\hat{S} - S_*\|_F^2 \end{aligned} \tag{49}$$

In the case when $\delta^2 < \varepsilon_F$, we have $k(\delta) \geq k(\sqrt{\bar{\varepsilon}}) = s$. In this case, we again use the fact that $\frac{\varphi(k)}{\lambda_k}$ is a nonincreasing function and the condition $\lambda_{k+1}/\lambda_k \leq c$ to show that

$$\begin{aligned} &\sqrt{\frac{\bar{t}}{nm}} \|\hat{S} - S_*\|_F \sqrt{\varphi(k(\delta)+1)} \\ &= \sqrt{\frac{\bar{t}}{mn}} \|W^{1/2}(\hat{S} - S_*)\|_F \sqrt{\delta^2 \varphi(k(\delta)+1)} \\ &\leq \sqrt{\frac{\bar{t}}{mn}} \|W^{1/2}(\hat{S} - S_*)\|_F \sqrt{\frac{\varphi(k(\delta)+1)}{\lambda_{k(\delta)}}} \\ &\leq \sqrt{c} \sqrt{\frac{\bar{t}}{mn}} \|W^{1/2}(\hat{S} - S_*)\|_F \sqrt{\frac{\varphi(k(\delta)+1)}{\lambda_{k(\delta)+1}}} \\ &\leq \sqrt{c} \sqrt{\frac{\bar{t}}{mn}} \|W^{1/2}(\hat{S} - S_*)\|_F \sqrt{\frac{\varphi(s+1)}{\lambda_{s+1}}} \\ &\leq \sqrt{c} \sqrt{\frac{\bar{t}}{mn}} \sqrt{\varepsilon_F} \|W^{1/2}(\hat{S} - S_*)\|_F \sqrt{\varphi(s+1)}. \end{aligned}$$

This allows us to deduce from (48) that

$$\begin{aligned} &|\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| \\ &\leq \sqrt{c} C \sqrt{\frac{\bar{t}}{mn}} \sqrt{\varepsilon_F} \|W^{1/2}(\hat{S} - S_*)\|_F \sqrt{\varphi(s+1)} \\ &+ 4\sqrt{2} \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n} \|\hat{S} - S_*\|_F \\ &\leq cC^2 \frac{\varphi(s+1) m \bar{t}}{n} + \frac{\varepsilon_F}{4m^2} \|W^{1/2}(\hat{S} - S_*)\|_F^2 \\ &+ 32 \max_{v \in V} \|P_L e_v\|^2 \Big(\frac{m\bar{t}}{n} \Big)^2 + \frac{1}{4m^2} \|\hat{S} - S_*\|_F^2. \end{aligned} \tag{50}$$

It follows from bounds (49) and (50) that with probability at least $1 - e^{-t}$,

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| \le (2 \vee c) C^2 \frac{\varphi(s+1) m \bar{t}}{n} + 64 \max_{v \in V} \|P_L e_v\|^2 \Big(\frac{m\bar{t}}{n} \Big)^2 \\ + \frac{1}{4m^2} \|\hat{S} - S_*\|_F^2 + \frac{\varepsilon_F}{4m^2} \|W^{1/2}(\hat{S} - S_*)\|_F^2, \end{aligned} \tag{51}$$

provided that

$$\delta = \frac{\|\hat{S} - S_*\|_F}{\|W^{1/2}(\hat{S} - S_*)\|_F} \in [\delta_-, \delta_+]. \tag{52}$$

It remains now to substitute bounds (28), (30), (32) and (51) in bound (20) to get that with some constants $C > 0, C_1 > 0$ depending only on c and with probability at least $1 - 2e^{-t}$

$$\begin{aligned} \frac{1}{m^2} \|\hat{S} - S_*\|_F^2 \le C \frac{\varphi(s+1) m (\bar{t} + t_m)}{n} \\ + \frac{\varepsilon_F}{m^2} \|W^{1/2} S_*\|_F^2 + C_1 \max_{v \in V} \|P_L e_v\|^2 \Big(\frac{m\bar{t}}{n} \Big)^2, \end{aligned} \tag{53}$$

where $t_m := t + \log(2m)$.

We still have to choose the values of δ_-, δ_+ and to handle the case when

$$\delta = \frac{\|\hat{S} - S_*\|_F}{\|W^{1/2}(\hat{S} - S_*)\|_F} \notin [\delta_-, \delta_+]. \tag{54}$$

First note that, since the largest eigenvalue of W is λ_m and it is bounded from above by m^ζ, we have

$$\|W^{1/2}(\hat{S} - S_*)\|_F \le \sqrt{\lambda_m} \|\hat{S} - S_*\|_F \le m^{\zeta/2} \|\hat{S} - S_*\|_F.$$

Thus, $\delta \ge m^{-\zeta/2}$. Next note that

$$\|W^{1/2} S_*\|_F^2 \le m^\zeta \|S_*\|_F^2 \le m^\zeta m^2,$$

where we also took into account that the absolute values of the entries of S_* are bounded by 1. It now follows from (34) that, under the assumption $\frac{2mt_m}{n} \le 1$,

$$\begin{aligned} \frac{1}{m^2} \|\hat{S} - S_*\|_F^2 \le \frac{3}{2} r m^2 \varepsilon^2 + 2\varepsilon_F m^\zeta \\ \le 24 r m^2 \frac{t + \log(2m)}{nm} + 2 \frac{m^\zeta}{\lambda_s} \le 12m + 2m^{2\zeta} \le 14 m^{2\zeta}, \end{aligned}$$

which holds with probability at least $1-e^{-t}$. Therefore, as soon as $\|W^{1/2}(\hat{S}-S_*)\|_F \geq m^2 n^{-\zeta}$, we have $\delta \leq 4n^{\zeta}m^{\zeta}$.

We will now take $\delta_- := m^{-\zeta/2}, \delta_+ := 4n^{\zeta}m^{\zeta}$. Then, the only case when (54) can possibly hold is if $\|W^{1/2}(\hat{S}-S_*)\|_F \leq m^2 n^{-\zeta}$. In this case, we can set

$$\delta := \frac{n^{\zeta}}{m^2}\|\hat{S}-S_*\|_F \in [\delta_-, \delta_+]$$

and follow the proof of bound (51) replacing throughout the argument $\|W^{1/2}(\hat{S}-S_*)\|_F$ with $m^2 n^{-\zeta}$. This yields

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, \hat{S}-S_*\rangle| &\leq (2\vee c)C^2 \frac{\varphi(s+1)m\bar{t}}{n} \\ +64 \max_{v\in V}\|P_L e_v\|^2 \Big(\frac{m\bar{t}}{n}\Big)^2 &+ \frac{1}{4m^2}\|\hat{S}-S_*\|_F^2 + \frac{1}{4}\varepsilon_F n^{-2\zeta}. \end{aligned} \tag{55}$$

Bound (55) can be now used instead of (51) to prove that

$$\begin{aligned} \frac{1}{m^2}\|\hat{S}-S_*\|_F^2 &\leq C\frac{\varphi(s+1)m(\bar{t}+t_m)}{n} \\ +\frac{\varepsilon_F}{m^2}\|W^{1/2}S_*\|_F^2 + C_1 \max_{v\in V}\|P_L e_v\|^2 \Big(\frac{m\bar{t}}{n}\Big)^2 &+ \varepsilon_F n^{-2\zeta} \end{aligned} \tag{56}$$

with some constants $C, C_1 > 0$ depending only on c.

Clearly, we can assume that $C_1 \geq 1$ and $\bar{t} \geq 1$. Since $m \leq n^2$ (recall that we even assumed that $mt_{n,m} \leq 1$), $\zeta \geq 1$, $\max_{v\in V}\|P_L e_v\|^2 \geq \frac{r}{m}$ and $\varepsilon_F \leq \lambda_{k_0}^{-1} \leq m^{\zeta}$, it is easy to check that

$$C_1 \max_{v\in V}\|P_L e_v\|^2 \Big(\frac{m\bar{t}}{n}\Big)^2 \geq \frac{m}{n^2} \geq \frac{m^{\zeta}}{n^{2\zeta}} \geq \varepsilon_F n^{-2\zeta}.$$

Thus, the last term of bound (56) can be dropped (with a proper adjustment of constant C_1).

Note also that with our choice of δ_-, δ_+

$$\bar{t} = t + \log(\log_F(\delta_+/\delta_-)+2) \leq t + \log\log_F(16 n^{\zeta} m^{(3/2)\zeta})$$

and $\bar{t} + t_m \leq 2t_{n,m}$. It is now easy to conclude that, with some constants C, C_1 depending only on c and with probability at least $1-3e^{-t}$

$$\begin{aligned} &\frac{1}{m^2}\|\hat{S}-S_*\|_F^2 \\ &\leq C\frac{\varphi(s+1)mt_{n,m}}{n} + \frac{\varepsilon}{m^2}\|W^{1/2}S_*\|_F^2 + C_1 \max_{v\in V}\|P_L e_v\|^2 \Big(\frac{m\bar{t}}{n}\Big)^2. \end{aligned} \tag{57}$$

The probability bound $1 - 3e^{-t}$ can be rewritten as $1 - e^{-t}$ by changing the value of constants C, C_1. Also, by changing the notation $s + 1 \mapsto s$, bound (57) yields (17). This completes the proof of the theorem.

□

Chapter III

LOW RANK ESTIMATION OF SMOOTH KERNELS ON GRAPHS

A *recommender system* is a platform that seeks to predict the rating that a user would give to an item. There are two main approaches to design a recommender system: content-based or collaborative filtering. On one hand, *content-based filtering* utilizes characteristics of items to recommend new items with similar properties; while on the other hand, *collaborative filtering* exploits information about the past behavior or the opinions of an existing user community for predicting which items the current user of the system will most probably like or be interested in. In this chapter, we consider scenarios where a hybrid approach combining content-based and collaborative filtering could lead to more accurate predictions.

Content-based filtering recommends items based on a comparison between the content of the items and a user profile. The content of each item is represented as a set of descriptors or terms, for instance, words that occur in a document. The user profile is represented with the same terms and built up by analyzing the content of items which have been seen by the user. In other words, these algorithms try to recommend items that are similar to those that a user liked in the past. In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended. Although, we often use text tags to describe the similarity among items, we could use a weighted graph for that purpose. Likewise, we can describe profiles using a properly designed weighted graph [53, 7].

From the collaborative filtering perspective, we can post the recommender system problem as a matrix completion problem. In this case, we formulate the problem

as that of inferring the contents of a partially observed *utility matrix*: each row represents a user, each column represents an item, and entries in the matrix represent a given user's rating of a given item. Our goal is to infer the unknown entries in the matrix from the observed entries –of which there are typically very few. To make useful predictions within this setting, we assume that the preference function can be decomposed into a small number of factors, resulting in the search for a low-rank matrix which approximates the partially observed utility matrix.

Collaborative filtering can perform in situations where it is difficult to describe items' content. On the other hand, since collaborative filtering relies only on previous users' ratings to produce recommendations, it usually requires more data than content-based filtering. For instance, a collaborative filtering method cannot give any information about an item that no user has rated before. In a low-rank matrix completion scenario, this means that we cannot make any prediction about a column for which we have not observed any entry. Nevertheless, this situation is easily resolved in content-based filtering, since we can make a recommendation comparing the descriptions of item content.

In this chapter, we consider a hybrid scheme where we exploit users profile and content of items (as in content-based filtering), and previous users' rating to items (as in collaborative filtering). We assume that the profile information is given by a weighted graph $G_{\mathcal{U}} = (\mathcal{U}, \mathcal{A}_{\mathcal{U}})$ with vertex set $\mathcal{U}$ representing users and symmetric matrix $\mathcal{A}_{\mathcal{U}}$ of nonnegative weights representing relations between users. Likewise, the items' content is given by a weighted graph $G_{\mathcal{V}} = (\mathcal{V}, \mathcal{A}_{\mathcal{V}})$ with vertex set $\mathcal{V}$ representing items and a symmetric matrix $\mathcal{A}_{\mathcal{V}}$ of nonnegative weights representing relations between items. Previous users' ratings are given by an incomplete utility matrix indexed by $\mathcal{U}$ and $\mathcal{V}$. Our goal is to predict the blanks of the utility matrix. We will base our completion of the utility matrix in two heuristics: first, we assume that few characteristics determine what items a user likes; and second, we assume

that similar users are likely to give similar ratings to similar items. Due to the first assumption we are interested in finding a low-rank matrix, while due to the second one we are interested in finding a smooth matrix with respect to the graphs $G_{\mathcal{U}}$ and $G_{\mathcal{V}}$.

3.1 Modeling the problem

We are interested in the problem of estimating a "smooth" and low-rank matrix $M_* : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ indexed by two weighted graphs $G_{\mathcal{U}} = (\mathcal{U}, \mathcal{A}_{\mathcal{U}})$ and $G_{\mathcal{V}} = (\mathcal{V}, \mathcal{A}_{\mathcal{V}})$ of size $m_{\mathcal{U}} \in \mathbb{N}$ and $m_{\mathcal{V}} \in \mathbb{N}$ respectively. As explained in section 3.1.2, we measure smoothness with respect to the geometry on the graphs induced by their graph Laplacians. We base the estimation on a finite number of noisy linear measurements of M_*. For simplicity, assume that these are the measurements of randomly picked entries of the target matrix M_*, which is a standard sampling model in matrix completion. More precisely, let $(U_j, V_j, Y_j), j = 1, \dots, n$ be n independent copies of a random triple (U, V, Y), where U and V are independent random vertices sampled from the uniform distribution $\Pi_{\mathcal{U}}$ in $\mathcal{U}$ and $\Pi_{\mathcal{V}}$ in $\mathcal{V}$ respectively, and $Y \in \mathbb{R}$ is a "measurement" of the matrix M_* at a random location (U, V) in the sense that $\mathbb{E}(Y|U, V) = M_*(U, V)$. In what follows, we assume that, for some constant $a > 0$, $|Y| \leq a$ a.s., which implies that $|M_*(u, v)| \leq a$ for $u \in \mathcal{U}$ and $v \in \mathcal{V}$. The target matrix M_* has to be estimated based on its i.i.d. measurements $(U_j, V_j, Y_j), j = 1, \dots, n$. Although, we introduced the problem in the context of recommender systems, our main motivation is mostly theoretical: we would like to explore to which extent taking into account smoothness of the target kernel could improve the existing methods of low rank recovery.

3.1.1 Estimation problem in the trace regression model

We consider the problem of estimating a matrix $M_* \in \mathbb{R}^{\mathcal{U} \times \mathcal{V}}$ based on observations $(X_1, Y_1), \dots, (X_n, Y_n)$ be n independent random pairs, where each X_k, $k = 1, \dots, n$ is a random matrix distributed according to $\hat{\Pi}_k$, $k = 1, \dots, n$ and each Y_k, $k = 1, \dots, n$

satisfies the *trace regression model*

$$\mathbb{E}(Y_k|X_k) = \langle M_*, X_k \rangle, \quad k = 1, \ldots, n$$

In this context, we refer to the matrices X_k, $k = 1, \ldots, n$ as the *design matrices* and to the observations Y_k, $k = 1, \ldots, n$ as the *response variables*. For simplicity, we concentrate in the case where all the design matrices X_k, $k = 1, \ldots, n$ are identically distributed according to a distribution $\hat{\Pi}$. Let M and N be two arbitrary real-valued matrices with rows indexed by $\mathcal{U}$ and columns indexed by $\mathcal{V}$. The following design dependent inner product is related to the trace regression model

$$\langle M, N \rangle_{L_2(\hat{\Pi})} = \int_{\mathbb{R}^{\mathcal{V} \times \mathcal{U}}} \langle M, X \rangle \langle N, X \rangle d\hat{\Pi}(X) = \mathbb{E}\langle M, X \rangle \langle N, X \rangle$$

Similarly, we define a distribution dependent inner product that is naturally related to the uniform sampling model. For that purpose, let $\Pi_{\mathcal{U}} \otimes \Pi_{\mathcal{V}}$ be the product distribution of $\Pi_{\mathcal{U}}$ and $\Pi_{\mathcal{V}}$. By independence, the random pair (U, V) is distributed according to $\Pi_{\mathcal{U}} \otimes \Pi_{\mathcal{V}}$. We are interested in the following inner product,

$$\langle M, N \rangle_{L_2(\Pi_{\mathcal{U}} \otimes \Pi_{\mathcal{V}})} = \int_{\mathcal{V} \times \mathcal{U}} M(u, v) N(u, v) d\Pi_{\mathcal{U}} \otimes \Pi_{\mathcal{V}}(u, v) = \mathbb{E}\langle M, N \rangle$$

Let $\{e_u \in \mathbb{R}^{\mathcal{U}} : u \in \mathcal{U}\}$ and $\{e_v \in \mathbb{R}^{\mathcal{V}} : v \in \mathcal{V}\}$ be the canonical orthonormal basis of the spaces $\mathbb{R}^{\mathcal{U}}$ and $\mathbb{R}^{\mathcal{V}}$ respectively. Let us consider the case where $\hat{\Pi}$ is the uniform distribution over the natural basis $\{e_u \otimes e_v \in \mathbb{R}^{\mathcal{U} \times \mathcal{V}} : u \in \mathcal{U}, v \in \mathcal{V}\}$ for the space of matrices $\mathbb{R}^{\mathcal{U} \times \mathcal{V}}$. Picking a matrix X randomly according to $\hat{\Pi}$ is equivalent to picking independently $U \in \mathcal{U}$ and $V \in \mathcal{V}$ from the uniform distributions $\Pi_{\mathcal{U}}$ and $\Pi_{\mathcal{V}}$ respectively and then setting $X = e_U \otimes e_V$. In other words, for our purposes, the uniform sampling model over the vertices $\mathcal{V}$ and $\mathcal{U}$ is equivalent to the trace regression model when the design matrices X_k, $k = 1, \ldots, m$ are sampled from the uniform distribution $\hat{\Pi}$. This equivalence is reinforced by noticing that, in this case, the inner products $\langle \cdot, \cdot \rangle_{L_2(\Pi_{\mathcal{U}} \otimes \Pi_{\mathcal{V}})}$ and $\langle \cdot, \cdot \rangle_{L_2(\hat{\Pi})}$ are the same.

The corresponding L_2-norm is naturally related to our problem, and it will be used to measure the estimation error. Since $\Pi_{\mathcal{U}}$ and $\Pi_{\mathcal{V}}$ are uniform distributions over $\mathcal{U}$ and $\mathcal{V}$ respectively, we note that

$$\langle M, N\rangle_{L_2(\Pi_{\mathcal{U}}\otimes\Pi_{\mathcal{V}})} = \langle M, N\rangle_{L_2(\hat{\Pi})} = \frac{1}{m_{\mathcal{U}}m_{\mathcal{V}}}\langle M, N\rangle$$
$$\|M\|^2_{L_2(\Pi_{\mathcal{U}}\otimes\Pi_{\mathcal{V}})} = \|M\|^2_{L_2(\hat{\Pi})} = \frac{1}{m_{\mathcal{U}}m_{\mathcal{V}}}\|M\|^2_F$$

In what follows, it will be often more convenient to use these rescaled versions rather than the actual Frobenius norm or Hilbert-Schmidt inner product.

3.1.2 Characterizing smoothness

Given two weighted graphs $G_{\mathcal{U}} = (\mathcal{U}, \mathcal{A}_{\mathcal{U}})$ and $G_{\mathcal{V}} = (\mathcal{V}, \mathcal{A}_{\mathcal{V}})$ of size $m_{\mathcal{U}} \in \mathbb{N}$ and $m_{\mathcal{V}} \in \mathbb{N}$ respectively, we consider the space $\mathcal{M}_{G_{\mathcal{U}}\times G_{\mathcal{V}}}$ of real-valued matrices $M : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ indexed by the vertex sets $\mathcal{U}$ and $\mathcal{V}$. A weighted graph can be naturally endowed with a geometry using the graph Laplacian operator. The geometry on the graphs $G_{\mathcal{U}}$ and $G_{\mathcal{V}}$ allows us to define a concept of "smoothness" for matrices in $\mathcal{M}_{G_{\mathcal{U}}\times G_{\mathcal{V}}}$ via discrete Sobolev norms.

A weighted graph G is a pair $(\mathcal{V}, \mathcal{A})$ where $\mathcal{V}$ is an arbitrary set and $\mathcal{A}$ is a symmetric matrix with nonnegative entries indexed by $\mathcal{V}$. The elements of $\mathcal{V}$ are called vertices, and each pair of vertices v_1 and v_2 form an edge. For an edge $e = \{v_1, v_2\}$, we interpret the entry $\mathcal{A}(v_1, v_2)$ as the weight of e. For each vertex $v \in \mathcal{V}$, we define $\deg(v) := \sum_{v'\in\mathcal{V}} \mathcal{A}(v, v')$. We identify the space of real-valued functions over $\mathcal{V}$ with the euclidean space $\mathbb{R}^{\mathcal{V}}$ endowed with the standard inner product $\langle\cdot,\cdot\rangle$ and the euclidean norm $\|\cdot\|$. We characterize the smoothness a function $f : \mathcal{V} \to \mathbb{R}$ by its energy,

$$\mathscr{E}^2_G(f) = \sum_{u,v\in\mathcal{V}} \mathcal{A}(u, v)|f(u) - f(v)|^2$$

In order to study the energy function $\mathscr{E}_G$, we introduce the Laplacian Δ_G of G,

$$\Delta_G(u,v) := \begin{cases} \deg(u) & u = v \\ -\mathcal{A}(u,v) & u \neq v \end{cases}$$

As in the case of simple graphs, the Laplacian induces a positive semi-definite bilinear form that is related to the energy function defined above. In other words, the Laplacian induces a geometry on the graph that is compatible with our measure of energy. To be precise,

$$\langle f, g\rangle_{\Delta_G} := \langle f, \Delta_G g\rangle = \langle \Delta_G^{1/2} f, \Delta_G^{1/2} g\rangle$$

$$\|f\|^2_{\Delta_G} := \langle f, f\rangle_{\Delta_G} = \mathscr{E}^2_G(f)$$

We characterize the smoothness of a matrix $M \in \mathcal{M}_{G_\mathcal{U} \times G_\mathcal{V}}$ in terms of the following Sobolev type norm

$$\|\Delta^{1/2}_{G_\mathcal{U}} M\|^2_F + \|\Delta^{1/2}_{G_\mathcal{V}} M^T\|^2_F$$

Note that if M has singular value decomposition $M = \sum_{k=1}^r \mu_k(\mathfrak{u}_k \otimes \mathfrak{v}_k)$, then

$$\|\Delta^{1/2}_{G_\mathcal{U}} M\|^2_F + \|\Delta^{1/2}_{G_\mathcal{V}} M^T\|^2_F =$$

$$\sum_{i,j=1}^r \mu_i\mu_j\langle \Delta_{G_\mathcal{U}}(\mathfrak{u}_i \otimes \mathfrak{v}_i), \mathfrak{u}_j \otimes \mathfrak{v}_j\rangle + \sum_{i,j=1}^r \mu_i\mu_j\langle \Delta_{G_\mathcal{V}}(\mathfrak{v}_i \otimes \mathfrak{u}_i), \mathfrak{v}_j \otimes \mathfrak{u}_j\rangle$$

$$= \sum_{k=1}^r \mu_k^2\langle \Delta_{G_\mathcal{U}}\mathfrak{u}_k, \mathfrak{u}_k\rangle + \sum_{k=1}^r \mu_k^2\langle \Delta_{G_\mathcal{V}}\mathfrak{v}_k, \mathfrak{v}_k\rangle = \sum_{k=1} \mu_k^2(\mathscr{E}^2_{G_\mathcal{U}}(\mathfrak{u}_k) + \mathscr{E}^2_{G_\mathcal{V}}(\mathfrak{v}_k))$$

so, essentially, the smoothness of a matrix M depends on the energy of the singular functions $\mathfrak{u}_k$ on the graph $G_\mathcal{U}$ and the energy of the singular functions $\mathfrak{v}_k$ on the graph $G_\mathcal{V}$. In what follows, we will often use rescaled versions of Sobolev norms,

$$\|\Delta^{1/2}_{G_\mathcal{U}} f\|^2_{L_2(\Pi_\mathcal{U})} = \frac{1}{m_\mathcal{U}}\|\Delta_{G_\mathcal{U}} f\|^2, \quad f \in \mathbb{R}^\mathcal{U}$$

$$\|\Delta^{1/2}_{G_\mathcal{V}} g\|^2_{L_2(\Pi_\mathcal{V})} = \frac{1}{m_\mathcal{V}}\|\Delta_{G_\mathcal{V}} g\|^2, \quad g \in \mathbb{R}^\mathcal{U}$$

$$\|\Delta^{1/2}_{G_\mathcal{U}} M\|^2_{L_2(\Pi_\mathcal{U}\otimes\Pi_\mathcal{V})} + \|\Delta^{1/2}_{G_\mathcal{V}} M^T\|^2_{L_2(\Pi_\mathcal{V}\otimes\Pi_\mathcal{U})} = \frac{1}{m_\mathcal{U} m_\mathcal{V}}\left[\|\Delta^{1/2}_{G_\mathcal{U}} M\|^2_F + \|\Delta^{1/2}_{G_\mathcal{V}} M^T\|^2_F\right]$$

3.1.3 Reduction to symmetric kernels

For simplicity, during the rest of this chapter, we concentrate in the case of estimating a symmetric kernel S_* over a weighted graph $G_{\mathcal{W}} = (\mathcal{W}, \mathcal{A}_{\mathcal{W}})$ based on uniform sample of its entries. By concentrating in this case, we lose little generality since we can reduce a non-symmetric matrix recovery problem over two graphs to a symmetric kernel recovery problem over one graph. In this reduction, we map an arbitrary non-symmetric matrix to a symmetric kernel using hermitian dilation. Under this map, all the inner products and norms are equivalent up to constants. As a result, we are able to translate any lower and upper bound in our analysis to the most general case by changing constants.

Given two weighted graphs $G_{\mathcal{U}} = (\mathcal{U}, \mathcal{A}_{\mathcal{U}})$ and $G_{\mathcal{V}} = (\mathcal{V}, \mathcal{A}_{\mathcal{V}})$ of size $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$ respectively, we construct their union as the graph $G_{\mathcal{U}\sqcup\mathcal{V}} := (\mathcal{U}\sqcup\mathcal{V}, \mathcal{A}_{\mathcal{U}\sqcup\mathcal{V}})$ with vertex set $\mathcal{U}\sqcup\mathcal{V}$ formed by the disjoint union of $\mathcal{U}$ and $\mathcal{V}$, and weight matrix $\mathcal{A}_{\mathcal{U}\sqcup\mathcal{V}}$ given by

$$\mathcal{A}_{\mathcal{U}\sqcup\mathcal{V}} := \begin{pmatrix} \mathcal{A}_{\mathcal{U}} & O_{m_{\mathcal{U}},m_{\mathcal{V}}} \\ O_{m_{\mathcal{V}},m_{\mathcal{U}}} & \mathcal{A}_{\mathcal{V}} \end{pmatrix}$$

where, $O_{k,l}$ denotes the $k \times l$ zero matrix for any natural numbers k and l.

For a weighted graph $G = (\mathcal{W}, \mathcal{A})$, let $\mathcal{S}_{\mathcal{W}}$ be the space of symmetric kernels $S : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$, that is the space of functions satisfying $S(w, w') = S(w', w)$ for each $w, w' \in \mathcal{W}$. We often identify the space $\mathcal{S}_{\mathcal{W}}$ with the space of symmetric matrices on $\mathcal{W}$. We embed the space $\mathcal{M}_{G_{\mathcal{U}}\times G_{\mathcal{V}}}$ into the space $\mathcal{S}_{\mathcal{U}\sqcup\mathcal{V}}$ using hermitian dilation, that is, for a matrix $M \in \mathcal{M}_{G_{\mathcal{U}}\times G_{\mathcal{V}}}$ we associate a symmetric kernel $\mathcal{S}(M) \in \mathcal{S}_{\mathcal{U}\sqcup\mathcal{V}}$ in the following way

$$\mathcal{S}(M) = \begin{pmatrix} O_{m_{\mathcal{U}},m_{\mathcal{V}}} & M \\ M^T & O_{m_{\mathcal{V}},m_{\mathcal{U}}} \end{pmatrix}$$

The matrix norms of $\mathcal{S}(M)$ are related to the matrix norms of M in the following way,

$$\|\mathcal{S}(M)\| = \|M\|, \qquad \|\mathcal{S}(M)\|_F = \sqrt{2}\|M\|_F. \qquad \|\mathcal{S}(M)\|_* = 2\|M\|_*$$

Similarly, the Sobolev type norm for the symmetric kernel $\mathcal{S}(M)$ with respect to the graph $G_{\mathcal{U}\sqcup\mathcal{V}}$ is related to the Sobolev type norm of the matrix M with respect to the graphs $G_{\mathcal{U}}$ and $G_{\mathcal{V}}$, that is,

$$\|\Delta_{G_{\mathcal{U}\sqcup\mathcal{V}}}^{1/2}\mathcal{S}(M)\|_F^2 = \|\Delta_{G_{\mathcal{U}}}^{1/2}M\|_F^2 + \|\Delta_{G_{\mathcal{V}}}^{1/2}M^T\|_F^2$$

Let $\Pi_{\mathcal{U}\sqcup\mathcal{V}}$ be the uniform distribution on $\mathcal{U}\sqcup\mathcal{V}$. Let M and N be two matrices over the graphs $G_{\mathcal{U}}$ and $G_{\mathcal{V}}$. We have the following relation between the distribution dependent inner products,

$$\langle\mathcal{S}(M),\mathcal{S}(N)\rangle_{L_2(\Pi_{\mathcal{U}\sqcup\mathcal{V}}\otimes\Pi_{\mathcal{U}\sqcup\mathcal{V}})} = \frac{m_{\mathcal{U}}m_{\mathcal{V}}}{(m_{\mathcal{U}}+m_{\mathcal{V}})^2}\langle M,N\rangle_{L_2(\Pi_{\mathcal{U}}\otimes\Pi_{\mathcal{V}})}$$

which leads to the corresponding relation for the L_2-norm

$$\|\mathcal{S}(M)\|^2_{L_2(\Pi_{\mathcal{U}\sqcup\mathcal{V}}\otimes\Pi_{\mathcal{U}\sqcup\mathcal{V}})} = \frac{m_{\mathcal{U}}m_{\mathcal{V}}}{(m_{\mathcal{U}}+m_{\mathcal{V}})^2}\|M\|^2_{L_2(\Pi_{\mathcal{U}}\otimes\Pi_{\mathcal{V}})}.$$

3.2 Estimation on symmetric kernels

We consider the problem of estimating a symmetric kernel $S_* \in \mathcal{S}_{\mathcal{V}}$ over a weighted graph $G = (\mathcal{V},\mathcal{A})$ of size $m \in \mathbb{N}$. We base our estimate on a finite number of noisy linear measurements of S_*. To be precise, let $(U_1,V_1,Y_1),\ldots,(U_n,V_n,Y_n)$ be independent copies of a random triple (U,V,Y) where U and V are independent random vertices sampled from the uniform distribution Π in $\mathcal{V}$, and Y is a random variable satisfying $E(Y|U,V) = S_*(U,V)$.

Let $\Pi^2 := \Pi\otimes\Pi$ be the distribution of random couple (U,V). We use the distribution dependent norm $\|\cdot\|_{L_2(\Pi^2)}$ to measure the estimation error. Denote by $\langle\cdot,\cdot\rangle_{L_2(\Pi^2)}$ the corresponding inner product. Since Π is the uniform distribution in $\mathcal{V}$, $\|S\|^2_{L_2(\Pi^2)} = m^{-2}\|S\|_F^2$ and $\langle S_1,S_2\rangle_{L_2(\Pi^2)} = m^{-2}\langle S_1,S_2\rangle$.

We will denote by $\{e_v : v\in\mathcal{V}\}$ the canonical orthonormal basis of the space $\mathbb{R}^{\mathcal{V}}$. Based on this basis, one can construct matrices $E_{u,v} = E_{v,u} = \frac{1}{2}(e_u\otimes e_v + e_v\otimes e_u)$. If $v_1,\ldots,v_m$ is an arbitrary ordering of the vertices in $\mathcal{V}$, then $\{E_{v_j,v_j} : j = 1,\ldots,m\}\cup$

$\{\sqrt{2}E_{v_i,v_j} : 1 \le i < j \le m\}$ is an orthonormal basis of the space $\mathcal{S}_{\mathcal{V}}$ of symmetric matrices with Hilbert–Schmidt inner product.

In standard matrix completion problems, $\mathcal{V}$ is a finite set with no further structure (i.e., the set of edges of the graph or the weight matrix are not specified). In this problem, a matrix version of LASSO is based on a trade-off between fitting the target matrix to the data using least squares and minimizing the nuclear norm

$$\hat{S} := \operatorname*{argmin}_{S \in \mathcal{S}_{\mathcal{V}}} \left[\frac{1}{n} \sum_{j=1}^{n} \left(Y_j - S(U_j, V_j) \right)^2 + \varepsilon \|S\|_* \right]. \tag{58}$$

This method and its modifications have been studied by a number of authors [12, 55, 47, 40, 38]. The following low-rank oracle inequality was proved in [40] for a "linearized version" of the matrix LASSO estimator $\hat{S}$. Assume that, for some constant $a > 0$, $|Y| \le a$ a.s. Let $t > 0$ and suppose that

$$\varepsilon \ge 4a \left(\sqrt{\frac{t + \log(2m)}{nm}} \vee \frac{2(t + \log(2m))}{n} \right)$$

Then, there exists a constant $C > 0$ such that with probability at least $1 - e^{-t}$

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le \inf_{S \in \mathcal{S}_V} \left[\|S - S_*\|^2_{L_2(\Pi^2)} + Cm^2\varepsilon^2 \operatorname{rank}(S) \right].$$

The last bound was proved in [39] for the matrix LASSO estimator (58) itself in the case when the domain of optimization problem is $\{S \in \mathcal{S}_V : \max_{u,v \in V} |S(u,v)| \le a\}$.

Remember that the smoothness of a symmetric kernel $S \in \mathcal{S}_{\mathcal{V}}$ on a graph $G = (\mathcal{V}, \mathcal{A})$ is given by the Sobolev type norm $\|\Delta_G^{1/2} S\|_F^2$. We often use the distribution dependent version of that norm $\|\Delta_G^{1/2} S\|^2_{L_2(\Pi^2)} = m^{-2} \|\Delta_G^{1/2} S\|$. In our analysis, we consider an arbitrary positive semi-definite matrix instead of Δ_G. We do so to emphasize the fact that other interpretations of the problem are possible. The positive semidefinite matrix W is fixed throughout the paper, and its spectral properties are crucial in our analysis. Assume that W has the following spectral representation

$$W = \sum_{k=1}^{m} \lambda_k (\phi_k \otimes \phi_k),$$

where $0 \leq \lambda_1 \leq \cdots \leq \lambda_m$ are the eigenvalues repeated with their multiplicities, and $\phi_1, \ldots, \phi_m$ are the corresponding orthonormal eigenfunctions (of course, there is a multiple choice of ϕ_k in the case of repeated eigenvalues). Let $k_0 := \min\{k \leq m : \lambda_k > 0\}$. We will assume in what follows that, for some constant $c \geq 1$, $\lambda_{k+1} \leq c\lambda_k$ for all $k \geq k_0$. It will be also convenient to set $\lambda_k := +\infty$ for $k > m$.

Let $\rho := \|W^{1/2} S_*\|_{L_2(\Pi^2)}$ and $r := \mathrm{rank}(S_*)$. It is easy to show (see the proof of Theorem 3.5 below) that kernel S_* can be approximated by the following kernel:

$$S_{*,l} := \sum_{i,j=1}^{l} \langle S_* \phi_i, \phi_j \rangle (\phi_i \otimes \phi_j)$$

with the approximation error

$$\|S_* - S_{*,l}\|^2_{L_2(\Pi^2)} \leq \frac{2\rho^2}{\lambda_{l+1}}. \tag{59}$$

Note that the kernel $S_{*,l}$ can be viewed as an $l \times l$ matrix (represented in the basis of eigenfunctions $\{\phi_j\}$) and $\mathrm{rank}(S_{*,l}) \leq r \wedge l$, so, one needs approximately $(r \wedge l)l$ parameters to characterize such matrices. Thus, one can expect, that such a kernel can be estimated, based on n linear measurements, with the squared $L_2(\Pi^2)$-error of the order $\frac{a^2(r \wedge l)l}{n}$. Taking into account the bound on the approximation error (59) and optimizing with respect to $l = 1, \ldots, m$, it would be also natural to expect the following error rate in the problem of estimation of the target kernel S_*:

$$\min_{1 \leq l \leq m} \left[\frac{a^2(r \wedge l)l}{n} \vee \frac{\rho^2}{\lambda_{l+1}} \right]. \tag{60}$$

We will show that such a rate is attained (up to constants and log factors) for a version of least squares method with a nonconvex complexity penalty; see Section 3.5. This method is not computationally tractable, so, we also study another method, based on convex penalization with a combination of nuclear norm and squared Sobolev type norm, and show that the rates are attained for such a method, too, provided that the target matrix satisfies a version low coherence assumption with respect to the basis of eigenfunctions of W. More precisely, denote the range of S_* by $\mathrm{range}(S_*)$,

and by $P_{\text{range}(S_*)}$ the orthogonal projection to range(S_*); we will prove error bounds involving a *coherence function*,

$$\varphi(S_*;\lambda) := \sum_{\lambda_j \le \lambda} \left\langle P_{\text{range}(S_*)}, \phi_j \otimes \phi_j \right\rangle$$

that characterizes the relationship between the kernel W defining the smoothness and the target kernel S_*; see Section 3.6 for more details; see also [41] for similar results in the case of "linearized least squares" estimator with double penalization. Finally, we prove minimax lower bounds on the error rate that are roughly of the order

$$\max_{1 \le l \le m} \left[\frac{a^2(r \wedge l)l}{n} \wedge \frac{\rho^2}{\lambda_l} \right]$$

subject to some extra conditions and with additional terms; see Section 3.3. In typical situations, this expression is, up to a constant, of the same order as the upper bound (60). For instance, if $\lambda_l \asymp l^{2\beta}$ for some $\beta > 1/2$, then the minimax error rate of estimation of the target kernel S_* is of the order

$$\left(\left(\frac{a^2 \rho^{1/\beta} r}{n} \right)^{2\beta/(2\beta+1)} \wedge \left(\frac{a^2 \rho^{2/\beta}}{n} \right)^{\beta/(\beta+1)} \wedge \frac{a^2 r m}{n} \right) \vee \frac{a^2}{n}$$

up to log factors. When m is sufficiently large, the term $\frac{a^2 r m}{n}$ will be dropped from the minimum, and we end up with a nonparametric convergence rate controlled by the smoothness parameter β and the rank r of the target matrix S_* (the dependence on m in the first two terms of the minimum is only in the log factors).

3.3 Minimax lower bounds

In this section, we derive minimax lower bounds on the $L_2(\Pi^2)$-error of an arbitrary estimator $\hat{S}$ of the target kernel S_* under the assumptions that the response variable Y is bounded by a constant $a > 0$, the rank of S_* is bounded by $r \le m$ and its Sobolev norm $\|W^{1/2} S_*\|_{L_2(\Pi^2)}$ is bounded by $\rho > 0$. More precisely, given $r = 1, \ldots, m$ and $\rho > 0$, denote by $\mathcal{S}_{r,\rho}$ the set of all symmetric kernels $S : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ such that $\text{rank}(S) \le r$ and $\|W^{1/2} S\|_{L_2(\Pi^2)} \le \rho$.

Given r, ρ and $a > 0$, let $\mathcal{P}_{r,\rho,a}$ be the set of all probability distributions of (U, V, Y) such that (U, V) is uniformly distributed in $\mathcal{V} \times \mathcal{V}$, $|Y| \le a$ a.s. and $\mathbb{E}(Y|U, V) = S_*(U, V)$, where $S_* \in \mathcal{S}_{r,\rho}$. For $P \in \mathcal{P}_{r,\rho,a}$, denote $S_P(U, V) := \mathbb{E}_P(Y|U, V)$.

Recall that $\{\phi_j, j = 1, \dots, m\}$ are the eigenfunctions of W orthonormal in the space $(\mathbb{R}^{\mathcal{V}}, \langle \cdot, \cdot \rangle)$. Then $\bar{\phi}_j := \sqrt{m}\phi_j, j = 1, \dots, m$ are orthonormal in $L_2(\Pi)$. We measure the "density" of these eigenfunctions by the following constant

$$Q_p := \max_{1 \le j \le m} \|\bar{\phi}_j\|^2_{L_p(\Pi)}, \qquad p \ge 2$$

and its "sparsity" by the constant

$$d := \max_{v \in V} \operatorname{card}\bigl\{j : \phi_j(v) \ne 0\bigr\},$$

We will obtain minimax lower bounds for classes of distributions $\mathcal{P}_{r,\rho,a}$ in two different cases. In the first case, we assume that, for some (relatively large) value of $p \ge 2$, the quantity Q_p is not too large. Roughly speaking, it means that most of the components of vectors $\phi_j \in \mathbb{R}^{\mathcal{V}}$ are uniformly small, say, $\phi_j(v) \asymp m^{-1/2}, v \in \mathcal{V}, j = 1, \dots, m$. In other words, the $m \times m$ matrix $(\phi_j(v))_{j=1,\dots,m, v \in \mathcal{V}}$ is "dense," so we refer to this case as a *"dense case"*.

The opposite case occurs when the constant d is small. In that case, the matrix $(\phi_j(v))_{j=1,\dots,m, v \in \mathcal{V}}$ is *"sparse"*, and therefore we refer to this case as a *"sparse case"*. A typical example occurs when basis of eigenfunctions $\{\phi_j, j = 1, \dots, m\}$ coincides with the canonical basis $\{e_v : v \in \mathcal{V}\}$ of $\mathbb{R}^{\mathcal{V}}$ (then, $d = 1$).

Denote $l_0 := k_0 \wedge 32$. In the *dense case*, the following theorem holds.

Theorem 3.1. Define

$$\delta_n^{(1)}(r, \rho, a) := \max_{l_0 \le l \le m} \left[\frac{a^2(r \wedge l)l}{n} \wedge \frac{\rho^2}{\lambda_l} \wedge \frac{1}{p-1} \frac{1}{Q_p^2} \frac{a^2(r \wedge l)}{l} \frac{1}{m^{4/p}} \right].$$

There exist constants $c_1, c_2 > 0$ such that

$$\inf_{\hat{S}_n} \sup_{P \in \mathcal{P}_{r,\rho,a}} \mathbb{P}_P\bigl\{\|\hat{S}_n - S_P\|^2_{L_2(\Pi^2)} \ge c_1 \delta_n^{(1)}(r, \rho, a)\bigr\} \ge c_2,$$

where the infimum is taken over all the estimators $\hat{S}_n$ based on n i.i.d. copies of (U, V, Y).

In fact, it will follow from the proof that, if $\lambda_{k_0} \leq \frac{n\rho^2}{a^2(r\wedge k_0)k_0}$ (i.e., the smallest nonzero eigenvalue of W is not too large), then the maximum in the definition of $\delta_n^{(1)}(r,\rho,a)$ can be extended to all $l = 1, \dots, m$.

Corollary 3.2. Let

$$\delta_n^{(2)}(r,\rho,a) := \max_{l_0\leq l\leq m}\left[\frac{a^2(r\wedge l)l}{n}\wedge\frac{\rho^2}{\lambda_l}\wedge\frac{1}{Q^2_{\log m}}\frac{a^2(r\wedge l)}{l}\frac{1}{\log m}\right].$$

There exist constants $c_1, c_2 > 0$ such that

$$\inf_{\hat{S}_n}\sup_{P\in\mathcal{P}_{r,\rho,a}}\mathbb{P}_P\big\{\|\hat{S}_n - S_P\|^2_{L_2(\Pi^2)} \geq c_1\delta_n^{(2)}(r,\rho,a)\big\} \geq c_2.$$

Proof. Take $p = \log m$ in the statement of Theorem 3.1 and observe that $m^{4/p} = e^4$ and $\frac{1}{p-1} \geq \frac{1}{\log m}$. □

Remark 1. It is easy to check that $e^{-2}Q_\infty \leq Q_{\log m} \leq Q_\infty$.

It is obvious that one can replace the quantity $\delta_n^{(1)}(r,\rho,a)$ in Theorem 3.1 (or the quantity $\delta_n^{(2)}(r,\rho,a)$ in Corollary 3.2) by the following smaller quantity:

$$\delta_n^{(3)}(r,\rho,a) := \max_{l_0\leq l\leq L}\left[\frac{a^2(r\wedge l)l}{n}\wedge\frac{\rho^2}{\lambda_l}\right],$$

where $L := \left[\frac{1}{Q_p m^{2/p}}\sqrt{\frac{n}{p-1}}\right]\wedge m$. Moreover, denote

$$\bar{l} := \max\left\{l = l_0,\dots,m : (r\vee l)l\lambda_l \leq \frac{\rho^2 n}{a^2}\right\}.$$

It is straightforward to check that

$$\max_{l_0\leq l\leq m}\left[\frac{a^2(r\wedge l)l}{n}\wedge\frac{\rho^2}{\lambda_l}\right] = \frac{a^2(r\wedge\bar{l})\bar{l}}{n}\vee\frac{\rho^2}{\lambda_{\bar{l}+1}}$$

and, if $\bar{l} \leq L$, then $\delta_n^{(3)}(r,\rho,a) = \frac{a^2(r\wedge\bar{l})\bar{l}}{n}\vee\frac{\rho^2}{\lambda_{\bar{l}+1}}$.

Example. Suppose that, for some $\beta > 1/2$, $\lambda_l \asymp l^{2\beta}, l = 1, \dots, m$ (in particular, it means that $\lambda_l \neq 0$ and $l_0 = k_0 = 1$). Then, an easy computation shows that

$$\bar{l} = (\check{l} \wedge m) \vee 1, \qquad \check{l} \asymp \left(\frac{\rho^2}{a^2}\frac{n}{r}\right)^{1/(2\beta+1)} \wedge \left(\frac{\rho^2 n}{a^2}\right)^{1/(2\beta+2)}.$$

Let $p = \log m$ and take $L := \left[\frac{1}{e^2 Q_p}\sqrt{\frac{n}{\log(m/e)}}\right] \bigwedge m$.

The condition $\bar{l} \leq L$ is satisfied, for instance, when either

$$e^2 Q_p \sqrt{\log(m/e)}\left(\frac{\rho^2}{a^2 r}\right)^{1/(2\beta+1)} \leq c' n^{1/2 - 1/(2\beta+1)}$$

or

$$e^2 Q_p \sqrt{\log(m/e)}\left(\frac{\rho}{a}\right)^{1/(\beta+1)} \leq c' n^{1/2 - 1/(2\beta+2)}$$

where $c' > 0$ is a small enough constant (this, essentially, means that n is sufficiently large). Under either of these conditions, we get the following expression for a minimax lower bound:

$$\left(\left(\frac{a^2 \rho^{1/\beta} r}{n}\right)^{2\beta/(2\beta+1)} \wedge \left(\frac{a^2 \rho^{2/\beta}}{n}\right)^{\beta/(\beta+1)} \wedge \frac{a^2 r m}{n}\right) \vee \frac{a^2}{n}. \tag{61}$$

We now turn to the *sparse case*.

Theorem 3.3. Let

$$\delta_n^{(4)}(r, \rho, a) := \max_{l_0 \leq l \leq m}\left[\frac{a^2 (r \wedge l) l}{n} \wedge \frac{\rho^2}{\lambda_l} \wedge \frac{a^2}{d \log m}\frac{l^2}{m^2}\right].$$

There exist constants $c_1, c_2 > 0$ such that

$$\inf_{\hat{S}_n} \sup_{P \in \mathcal{P}_{r,\rho,a}} \mathbb{P}_P\left\{\|\hat{S}_n - S_P\|^2_{L_2(\Pi^2)} \geq c_1 \delta_n^{(4)}(r, \rho, a)\right\} \geq c_2.$$

It will be clear from the upper bounds of Section 3.5 (see the remark after Theorem 3.5) that, at least in a special case when $\{\phi_j\}$ coincides with the canonical basis of $\mathbb{R}^{\mathcal{V}}$, the additional term $\frac{a^2}{d \log m}\frac{l^2}{m^2}$ is correct (up to a log factor). At the same time, most likely, the "third terms" of the bounds of Theorem 3.1 (in the dense case) and Theorem 3.3 (in the sparse case) have not reached their final form yet. A more sophisticated construction of "well separated" subsets of $\mathcal{P}_{r,\rho,a}$ might be needed to achieve

this goal. The main difficulty in the proof given below is related to the fact that we have to impose constraints, on the one hand, on the entries of the target matrix represented in the canonical basis and, on the other hand, on the Soblolev type norm $\|W^{1/2}S\|_{L_2(\Pi^2)}$ (for which it is convenient to use the representation in the basis of eigenfunctions of W). Due to this fact, we are using the last representation in our construction, and we have to use an argument based on the properties of Rademacher sums to ensure that the entries of the matrix represented in the canonical basis are uniformly bounded by a. This is the reason why the "third terms" occur in the bounds of Theorems 3.1 and 3.3. In this case, when the constraints are only on the norm $\|W^{1/2}S\|_{L_2(\Pi^2)}$ and on the variance of the noise and there are no constraints on $\|S\|_{L_\infty}$, it is much easier to prove the lower bound of the order $\max_{l_0\le l\le m}[\frac{\sigma^2(r\wedge l)l}{n}\wedge\frac{\rho^2}{\lambda_l}]$ without any additional terms. Note, however, that the condition $\|S_*\|_{L_\infty}\le a$ is of importance in the following sections to obtain the upper bounds for penalized least squares estimators that match the lower bounds up to log factors.

3.4 Proof of lower bounds

Proof of Theorem 3.1. The proof relies on several well-known facts stated below. In what follows, $K(\mu\|\nu) := -\mathbb{E}_\mu \log\frac{d\nu}{d\mu}$ denotes *Kullback-Leibler divergence* between two probability measures μ,ν defined on the same space and such that ν is absolutely continuous with respect to μ (denoted by $\nu\ll\mu$). We will denote by $P^{\otimes n}$ the n-fold product measure $P^{\otimes n} := P\otimes P\cdots\otimes P$. The following proposition is a version of Theorem 2.5 in [59].

Proposition 3.4. Let $\mathcal{P}$ be a finite set of distributions of (U,V,Y) such that the following assumptions hold:

1. there exists $P_0\in\mathcal{P}$ such that for all $P\in\mathcal{P}$, $P\ll P_0$;

2. there exists $\alpha \in (0, 1/8)$ such that

$$\sum_{P \in \mathcal{P}} K\big(P_0^{\otimes n} \| P^{\otimes n}\big) \le \alpha\big(\mathrm{card}(\mathcal{P}) - 1\big) \log\big(\mathrm{card}(\mathcal{P}) - 1\big);$$

3. for all $P_1, P_2 \in \mathcal{P}$, $\|S_{P_1} - S_{P_2}\|^2_{L_2(\Pi^2)} \ge 4s^2 > 0$.

Then, there exists a constant $\beta > 0$ such that

$$\inf_{\hat{S}_n} \max_{P \in \mathcal{P}} \mathbb{P}_P\big\{\|\hat{S}_n - S_P\|^2_{L_2(\Pi^2)} \ge s^2\big\} \ge \beta > 0. \tag{62}$$

We will also use Varshamov–Gilbert bound (see [59], Lemma 2.9, page 104), Sauer's lemma (see [38], page 39) and the following elementary bound for Rademacher sums ([17], page 21): for all $p \ge 2$,

$$\mathbb{E}^{1/p}\left|\sum_{j=1}^N \varepsilon_j t_j\right|^p \le \sqrt{p-1}\left(\sum_{j=1}^N t_j^2\right)^{1/2}, \qquad (t_1, \dots, t_N) \in \mathbb{R}^N, \tag{63}$$

where $\varepsilon_1, \dots, \varepsilon_N$ are i.i.d. Rademacher random variables (i.e., $\varepsilon_j = +1$ with probability $1/2$ and $\varepsilon_j = -1$ with the same probability).

We will start the proof with constructing a "well separated" subset $\mathcal{P}$ of the class of distributions $\mathcal{P}_{r,\rho,a}$ that will allow us to use Proposition 3.4. Fix $l \le m$, $l \ge 32$ and $\kappa > 0$. Denote $l' = [l/2], l'' = l - l'$. First assume that $r \le l''$. Denote $R_\sigma := \kappa((\sigma_{ij}) : i = 1, \dots, l', j = 1, \dots, r)$, where $\sigma_{ij} = +1$ or $\sigma_{ij} = -1$. Let $\mathcal{R}_{l',r} = \{R_\sigma : \sigma \in \{-1, 1\}^{l' \times r}\}$ (so, $\mathcal{R}_{l',r}$ is the class of all $l' \times r$ matrices with entries $+\kappa$ or $-\kappa$). Given $R \in \mathcal{R}_{l',r}$, let

$$\tilde{R} := \begin{pmatrix} R & R & \cdots & R & O_{l',l^*} \end{pmatrix}$$

be the $l' \times l''$ matrix that consists of $[l''/r]$ blocks R and the last block O_{l',l^*}, where $l^* := l'' - [l''/r]r$ and O_{k_1,k_2} is the $k_1 \times k_2$ zero matrix. Finally, define the following symmetric $m \times m$ matrix:

$$R^\diamondsuit := \begin{pmatrix} O_{l',l'} & \tilde{R} & O_{l',m-l} \\ \tilde{R}^T & O_{l'',l''} & O_{l'',m-l} \\ O_{m-l,l'} & O_{m-l,l''} & O_{m-l,m-l} \end{pmatrix}$$

Now, given $\sigma \in \{-1,1\}^{l' \times r}$, define a symmetric kernel $K_\sigma : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$,

$$K_\sigma := \sum_{i,j=1}^{m} \left(R_\sigma^{\diamondsuit}\right)_{ij} (\phi_i \otimes \phi_j).$$

It is easy to see that

$$\begin{aligned} K_\sigma(u,v) &= K'_\sigma(u,v) + K'_\sigma(v,u), \\ K'_\sigma(u,v) &= \kappa \sum_{i=1}^{l'} \sum_{j=1}^{r} \sigma_{ij} \phi_i(u) \sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}(v). \end{aligned} \tag{64}$$

Let $\Lambda := \{\sigma \in \{-1,1\}^{l' \times r} : \max_{u,v \in V} |K_\sigma(u,v)| \le a\}$. We will show that, if κ is sufficiently small (its precise value to be specified later), then the set Λ contains at least three quarters of the points of the combinatorial cube $\{-1,1\}^{l' \times r}$. To this end, define $\xi := \max_{u,v \in V} |K_\varepsilon(u,v)|$, where $\varepsilon \in \{-1,1\}^{l' \times r}$ is a random vector with i.i.d. Rademacher components. Assume, in addition, that ε and (U,V) are independent. It is enough to show that $\xi \le a$ with probability at least $3/4$. We have

$$\begin{aligned} \mathbb{P}\{\xi \ge a\} &\le \sum_{u,v \in V} \mathbb{P}\Big\{\Big|K_\varepsilon(u,v)\Big| \ge a\Big\} \\ &= m^2 \mathbb{E}\mathbb{P}\Big\{\Big|K_\varepsilon\big(U,V\big)\Big| \ge a | U, V\Big\} \\ &= m^2 \mathbb{P}\Big\{\Big|K_\varepsilon\big(U,V\big)\Big| \ge a\Big\} \le \frac{m^2 \mathbb{E}|K_\varepsilon(U,V)|^p}{a^p}. \end{aligned}$$

We will use bound (63) to control $\mathbb{E}(|K_\varepsilon(U,V)|^p | U,V)$ (recall that $K_\varepsilon(u,v)$, $u,v \in V$ is a Rademacher sum). Denote

$$\tau^2(u,v) := \sum_{i=1}^{l'} \sum_{j=1}^{r} \phi_i^2(u) \left(\sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}(v) \right)^2.$$

Observe that $\tau^2(u,v) \le \frac{l''}{r} q(l',u) q(l'',v) \le q(l,u) q(l,v) \frac{l}{r}$, where $q(l,u) := \sum_{j=1}^{l} \phi_j^2(u), u \in V$, and we used the bound

$$\left(\sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}(v) \right)^2 \le \frac{l''}{r} \sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}^2(v). \tag{65}$$

Thus, applying (63) to the Rademacher sum K'_ε, we get

$$\begin{aligned}\mathbb{E}\Big|K_\varepsilon(u,v)\Big|^p &\le 2^{p-1}\big(\mathbb{E}\Big|K'_\varepsilon(u,v)\Big|^p + \mathbb{E}\Big|K'_\varepsilon(v,u)\Big|^p\big)\\ &\le 2^p(p-1)^{p/2}\kappa^p\big(\tau^2(u,v)\vee\tau^2(v,u)\big)^{p/2}\\ &\le 2^p(p-1)^{p/2}\kappa^p q^{p/2}(l,u)q^{p/2}(l,v)\left(\frac{l}{r}\right)^{p/2}.\end{aligned}$$

Given $p\in[2,+\infty]$, denote $Q_p(l) := \|\frac{m}{l}q(l,\cdot)\|_{L_{p/2}(\Pi)} = \|\frac{1}{l}\sum_{j=1}^l \bar\phi_j^2\|_{L_{p/2}(\Pi)}$ for $l = 1,\dots,m$. This yields

$$\begin{aligned}\mathbb{E}\Big|K_\varepsilon\big(U,V\big)\Big|^p &= \mathbb{E}\mathbb{E}\big(\Big|K_\varepsilon\big(U,V\big)\Big|^p|U,V\big)\\ &\le 2^p(p-1)^{p/2}\kappa^p\left(\frac{l}{r}\right)^{p/2}\mathbb{E}\big(q^{p/2}(l,U)q^{p/2}\big(l,V\big)\big)\\ &= 2^p(p-1)^{p/2}\kappa^p\left(\frac{l}{r}\right)^{p/2}\big(\mathbb{E}q^{p/2}(l,V)\big)^2\\ &= 2^p(p-1)^{p/2}\kappa^p\left(\frac{l}{r}\right)^{p/2}\left(\frac{l}{m}\right)^p Q_p^p(l).\end{aligned}$$

Substituting the last bound into (3.4), we get

$$\mathbb{P}\{\xi\ge a\}\le\frac{m^2\mathbb{E}|K_\varepsilon(U,V)|^p}{a^p}\le m^2 2^p(p-1)^{p/2}\frac{\kappa^p}{a^p}\left(\frac{l}{r}\right)^{p/2}\left(\frac{l}{m}\right)^p Q_p^p(l).$$

Now, to get $\mathbb{P}\{\xi\ge a\}\le 1/4$, it is enough to take

$$\kappa\le 2^{-(1+2/p)}(p-1)^{-1/2}\frac{1}{Q_p(l)}\frac{m}{l}\frac{a\sqrt{r}}{\sqrt{l}}\frac{1}{m^{2/p}}. \tag{66}$$

Next observe that

$$|\Lambda|\ge\frac{3}{4}2^{l'r} > \sum_{k=0}^{[l'r/2]}\binom{l'r}{k}$$

It follows from Sauer's lemma that there exists a subset $J\subset\{(i,j):1\le i\le l', 1\le j\le r\}$ with $|J| = [l'r/2]+1$ and such that $\pi_J(\Lambda) = \{-1,1\}^J$, where $\pi_J:\{-1,1\}^{l'\times r}\to\{-1,1\}^J$ is the projection:

$$\pi_J(\sigma_{ij}: i=1,\dots,l', j=1,\dots,r) = (\sigma_{ij}:(i,j)\in J).$$

Since $l \geq 32$, we have $l'r \geq 16$ and $|J| \geq 8$. We can now apply Varshamov–Gilbert bound to the combinatorial cube $\{-1,1\}^J$ to prove that there exists a subset $E \subset \{-1,1\}^J$ such that $|E| \geq 2^{l'r/16}+1$ and, for all $\sigma', \sigma'' \in E, \sigma' \neq \sigma''$,

$$\sum_{(i,j)\in J} I(\sigma'_{ij} \neq \sigma''_{ij}) \geq \frac{l'r}{16}.$$

It is now possible to choose a subset Λ' of Λ such that $|\Lambda'| = |E|$ and $\pi_J(\Lambda') = E$. Then, we have $|\Lambda'| \geq 2^{l'r/16}+1$ and

$$\sum_{i=1}^{l'} \sum_{j=1}^{r} I\big(\sigma'_{ij} \neq \sigma''_{ij}\big) \geq \frac{l'r}{16} \tag{67}$$

for all $\sigma', \sigma'' \in \Lambda', \sigma' \neq \sigma''$.

We are now in a position to define the set of distributions $\mathcal{P}$. For $\sigma \in \Lambda'$, denote by P_σ the distribution of (U,V,Y) such that (U,V) is uniform in $\mathcal{V} \times \mathcal{V}$ and the conditional distribution of Y given (U,V) is defined as follows:

$$\mathbb{P}_{P_\sigma}\big\{Y = \delta a | U,V\big\} = p_\sigma\big(U,V\big) = 1/2 + \delta K_\sigma\big(U,V\big)/8a, \qquad \delta \in \{-1,+1\}.$$

Since $|K_\sigma(U,V)| \leq a$ for all $\sigma \in \Lambda'$, we have $p_\sigma(U,V) \in [3/8, 5/8], \sigma \in \Lambda$. Denote $\mathcal{P} := \{P_\sigma : \sigma \in \Lambda'\}$. For $P = P_\sigma \in \mathcal{P}$, we have

$$S_P(u,v) = \mathbb{E}\big(Y | X = u, X' = v\big) = \tfrac{1}{4} K_\sigma(u,v).$$

Note that $\operatorname{rank}(S_P) = \operatorname{rank}(K_\sigma) = \operatorname{rank}(R_\sigma^\diamondsuit) \leq r$; see the definitions of K_σ and $R_\sigma^\diamondsuit$. Moreover, we have

$$\big\|W^{1/2} K_\sigma\big\|_F^2 = \Big\|W^{1/2} \sum_{i,j=1}^{m} \big(R_\sigma^\diamondsuit\big)_{ij} (\phi_i \otimes \phi_j)\Big\|_F^2 = \sum_{i,j=1}^{l} \lambda_i \big(R_\sigma^\diamondsuit\big)_{ij}^2 \leq \lambda_l \|K_\sigma\|_F^2$$

and

$$\begin{aligned} &\|K_\sigma\|_F^2 \\ &= \Big\| \kappa \sum_{i=1}^{l'} \sum_{j=1}^{r} \sigma_{ij} \sum_{k=0}^{[l''/r]-1} \phi_i \otimes \phi_{l'+rk+j} + \kappa \sum_{i=1}^{r} \sum_{j=1}^{l'} \sigma_{ji} \sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+i} \otimes \phi_j \Big\|_F^2 \\ &\leq 2\kappa^2 l' r \big[l''/r\big] \leq \kappa^2 l^2. \end{aligned}$$

Therefore, $\|W^{1/2}K_\sigma\|^2_{L_2(\Pi^2)} \le \lambda_l \kappa^2 \frac{l^2}{m^2}$, so, we have

$$\left\|W^{1/2}S_{P_\sigma}\right\| = \tfrac{1}{16}\left\|W^{1/2}K_\sigma\right\|^2_{L_2(\Pi^2)} \le \rho^2, \tag{68}$$

provided that

$$\kappa \le \frac{m}{l}\frac{4\rho}{\sqrt{\lambda_l}}. \tag{69}$$

We can conclude that, for all $P \in \mathcal{P}$, $S_P \in \mathcal{S}_{r,\rho}$ provided that κ satisfies conditions (66) and (69). Since also $|Y| \le a$, we have that $\mathcal{P} \subset \mathcal{P}_{r,\rho,a}$.

Next, we check that $\mathcal{P}$ satisfies the conditions of Proposition 3.4. It is easy to see that, for all $\sigma, \sigma' \in \Lambda' P_{\sigma'} \ll P_\sigma$ and

$$\begin{aligned} &K(P_\sigma \| P_{\sigma'}) \\ &= \mathbb{E}\Big(p_\sigma\big(U,V\big)\log\frac{p_\sigma(U,V)}{p_{\sigma'}(U,V)} + \big(1-p_\sigma\big(U,V\big)\big)\log\frac{1-p_\sigma(U,V)}{1-p_{\sigma'}(U,V)}\Big). \end{aligned}$$

Using the elementary inequality $-\log(1+u) \le -u+u^2, |u| \le 1/2$ and the fact that $p_\sigma(U,V) \in [3/8, 5/8], \sigma \in \Lambda$, we get that

$$K(P_\sigma \| P_{\sigma'}) \le \frac{6}{8^2 a^2}\|K_\sigma - K_{\sigma'}\|_{L_2(\Pi^2)} \le \frac{1}{10a^2m^2}\|K_\sigma - K_{\sigma'}\|_F^2, \sigma, \qquad \sigma' \in \Lambda'.$$

A simple computation based on the definition of $K_\sigma, K_{\sigma'}$ easily yields that

$$\|K_\sigma - K_{\sigma'}\|_F^2 \le 8\kappa^2 l' r\big[l''/r\big] \le 8\kappa^2 l' l'' \le 4\kappa^2 l^2.$$

Thus, for the n-fold product-measures $P_\sigma^{\otimes n}, P_{\sigma'}^{\otimes n}$, we get

$$K\big(P_\sigma^{\otimes n} \| P_{\sigma'}^{\otimes n}\big) = nK(P_\sigma \| P_{\sigma'}) \le \frac{4n\kappa^2}{10a^2}\frac{l^2}{m^2}.$$

For a fixed $\sigma \in \Lambda'$, this yields

$$\begin{aligned} &\frac{1}{|\Lambda'|-1}\sum_{\sigma' \in \Lambda'} K\big(P_\sigma^{\otimes n} \| P_{\sigma'}^{\otimes n}\big) \\ &\le \frac{4n\kappa^2}{10a^2}\frac{l^2}{m^2} \le \frac{1}{10}\frac{l'r}{16} \le \frac{1}{10}\log\big(|\Lambda'|-1\big), \end{aligned} \tag{70}$$

provided that

$$\kappa \le \frac{1}{16}a\frac{m}{l}\sqrt{\frac{rl}{n}}. \tag{71}$$

It remains to use (67) and the definition of kernels K_σ to bound from below the squared distance $\|K_\sigma - K_{\sigma'}\|^2_{L_2(\Pi^2)}$ for $\sigma, \sigma' \in \Lambda', \sigma \neq \sigma'$,

$$\|K_\sigma - K_{\sigma'}\|^2_{L_2(\Pi^2)} = m^{-2}\|K_\sigma - K_{\sigma'}\|^2_F \geq 4m^{-2}\kappa^2 \frac{l'r}{16}\big[l''/r\big] \geq \frac{1}{64}\kappa^2\frac{l^2}{m^2}.$$

Since $S_{P_\sigma} = \frac{1}{4}K_\sigma$, this implies that

$$\|S_P - S_{P'}\|^2_{L_2(\Pi^2)} \geq 2^{-10}\kappa^2\frac{l^2}{m^2}, \qquad P, P' \in \mathcal{P}, P \neq P'. \tag{72}$$

In view of (66), (71) and (69), we now take

$$\kappa := \frac{1}{16}a\frac{m}{l}\sqrt{\frac{rl}{n}} \wedge \frac{m}{l}\frac{4\rho}{\sqrt{\lambda_l}} \wedge 2^{-(1+2/p)}(p-1)^{-1/2}\frac{1}{Q_p(l)}\frac{m}{l}\frac{a\sqrt{r}}{\sqrt{l}}\frac{1}{m^{2/p}}.$$

With this choice of κ, $\mathcal{P} := \{P_\sigma : \sigma \in \Lambda'\} \subset \mathcal{P}_{r,a,\rho}$. In view of (72) and (70), we can use Proposition 3.4 to get

$$\begin{aligned} &\inf_{\hat{S}} \sup_{P\in\mathcal{P}_{r,a,\rho}} \mathbb{P}_P\big\{\|\hat{S} - S_P\|^2_{L_2(\Pi^2)} \geq c_1\delta_n\big\} \\ &\geq \inf_{\hat{S}} \sup_{P\in\mathcal{P}} \mathbb{P}_P\big\{\|\hat{S} - S_P\|^2_{L_2(\Pi^2)} \geq c_1\delta_n\big\} \geq c_2, \end{aligned} \tag{73}$$

where $\delta_n := \frac{a^2rl}{n} \wedge \frac{\rho^2}{\lambda_l} \wedge \frac{1}{p-1}\frac{1}{Q_p^2(l)}\frac{a^2r}{l}\frac{1}{m^{4/p}}$ and $c_1, c_2 > 0$ are constants.

In the case when $r > l''$, bound (73) still holds with

$$\delta_n := \frac{a^2l^2}{n} \wedge \frac{\rho^2}{\lambda_l} \wedge \frac{1}{p-1}\frac{a^2}{Q_p^2(l)}\frac{1}{m^{4/p}}.$$

The proof is an easy modification of the argument in the case when $r \leq l''$. For $r > l''$, the construction becomes simpler: namely, we define

$$R^\flat := \begin{pmatrix} O_{l',l'} & R & O_{l',m-l} \\ R^T & O_{l'',l''} & O_{l'',m-l} \\ O_{m-l,l'} & O_{m-l,l''} & O_{m-l,m-l} \end{pmatrix}$$

where $R \in \mathcal{R}_{l',l''}$, and, based on this, redefine kernels $K_\sigma, \sigma \in \{-1,1\}^{l'\times l''}$. The proof then goes through with minor simplifications.

Thus, in both cases $r > l''$ and $r \le l''$, (73) holds with

$$\delta_n = \delta_n(l) := \frac{a^2(r \wedge l)l}{n} \wedge \frac{\rho^2}{\lambda_l} \wedge \frac{1}{p-1}\frac{1}{Q_p^2(l)}\frac{a^2(r \wedge l)}{l}\frac{1}{m^{4/p}}.$$

This is true under the assumption that $l \ge 32$. Note also that

$$Q_p(l) \le \max_{1 \le j \le m} \|\bar{\phi}_j\|^2_{L_p(\Pi)} = Q_p$$

Thus, we can replace $Q_p^2(l)$ by the upper bound Q_p^2 in the definition of $\delta_n(l)$.

We can now choose $l \in \{32, \dots, m\}$ that maximizes $\delta_n(l)$ to get bound (73) with $\delta_n := \min_{32 \le l \le m} \delta_n(l)$. This completes the proof in the case when $k_0 \ge 32$ and $l_0 = 32$. If $k_0 < 32$, it is easy to use the condition $\lambda_{l+1} \le c\lambda_l, l \ge k_0$ and to show that $\min_{32 \le l \le m} \delta_n(l) \le c' \min_{k_0 \le l \le m} \delta_n(l)$, where c' is a constant depending only on c. This completes the proof in the remaining case. □

Proof of Theorem 3.3. The only modification of the previous proof is to replace bound (65) by

$$\left(\sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}(v)\right)^2 \le d \sum_{k=0}^{[l''/r]-1} \phi^2_{l'+rk+j}(v).$$

Then, the outcome of the next several lines of the proof is that $\mathbb{P}\{\xi \ge a\} \le 1/4$ provided that (instead of (66))

$$\kappa \le 2^{-(1+2/p)}(p-1)^{-1/2}\frac{1}{Q_p(l)}\frac{m}{l}\frac{a}{\sqrt{d}}\frac{1}{m^{2/p}}.$$

As a result, at the end of the proof, we get that (73) holds with

$$\delta_n = \delta_n(l) := \frac{a^2(r \wedge l)l}{n} \wedge \frac{\rho^2}{\lambda_l} \wedge \frac{1}{p-1}\frac{1}{Q_p^2(l)}\frac{a^2}{d}\frac{1}{m^{4/p}}.$$

It remains to observe that $Q_p(l) \le \frac{m}{l}$, which follows from the fact that

$$\sum_{j=1}^{l} \phi_j^2(v) = \sum_{j=1}^{l} \langle \phi_j, e_v \rangle^2 \le \sum_{j=1}^{m} \langle \phi_j, e_v \rangle^2 = 1, \qquad v \in V,$$

and to take $p = \log m$ to complete the proof. □

3.5 Least squares estimators with nonconvex penalties

In this section, we derive upper bounds on the squared $L_2(\Pi^2)$-error for a least squares estimator of the target kernel S_* with a non-convex feasible region. An appropriate choice of the feasible region will allow us to prove upper bounds matching the lower bounds up to log factors in some cases of interest. Firstly, we pick the feasible region assuming some information about the target matrix S_* (like its rank and how well it can be approximated by a small number of eigenvectors of W). Secondly, we consider a procedure to pick the unknown parameters adaptively.

In order to define such non-convex feasible region of interest, we introduce some subsets of symmetric kernels. For a kernel $S \in \mathcal{S}_{\mathcal{V}}$, let S^a denote the clipping of S by a. That is, $S^a(u,v) = S(u,v)$ if $|S(u,v)| \leq a$, $S^a(u,v) = a$ if $S(u,v) > a$ and $S^a(u,v) = -a$ if $S(u,v) < -a$, for each u and v in $\mathcal{V}$. Let $\mathcal{S}_r(l;a)$ be the set of symmetric kernels on $\mathcal{V}$ of rank at most r, $L_2(\Pi^2)$-norm bounded by a and with range in the linear span of $\{\phi_1, \ldots, \phi_l\}$. To be precise,

$$\mathcal{S}_r(l;a) := \left\{ S \in \mathcal{S}_V : \operatorname{rank}(S) \leq r, \|S\|_{L_2(\Pi^2)} \leq a, S = \sum_{i,j=1}^{l} s_{ij}(\phi_i \otimes \phi_j) \right\}.$$

Lastly, we define the set $\mathcal{S}_r(l;a)$ of clipped matrices from $\mathcal{S}_r(l,a)$,

$$\bar{\mathcal{S}}_r(l;a) := \{S^a : S \in \mathcal{S}_r(l;a)\}$$

Note that the sets $\mathcal{S}_r(l;a)$ and $\bar{\mathcal{S}}_r(l;a)$ are not convex.

3.5.1 Least square estimator

We are interested in the following least squares estimator of the target matrix S_*:

$$\hat{S}_l := \hat{S}_{r,l,a} := \operatorname*{argmin}_{S \in \bar{\mathcal{S}}_r(l;a)} \frac{1}{n} \sum_{j=1}^{n} \big(Y_j - S\big(U_j, V_j\big)\big)^2, \tag{74}$$

where l and r are parameters that we will choose adaptively as explained below. Note that the optimization problem (74) is not convex. We will prove the following result under the assumption that $|Y| \leq a$ a.s. Recall the definition of the class of kernels $\mathcal{S}_{r,\rho}$ in Section 3.3.

Theorem 3.5. There exist constants $C > 0, A > 0$ such that, for all $t > 0$, with probability at least $1 - e^{-t}$,

$$\begin{aligned} &\|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \\ &\leq 2 \inf_{S \in \bar{\mathcal{S}}_r(l;a)} \|S - S_*\|^2_{L_2(\Pi^2)} + C\biggl(\frac{a^2(r \wedge l)l}{n} \log\biggl(\frac{Anm}{(r \wedge l)l}\biggr) + \frac{a^2 t}{n}\biggr). \end{aligned} \tag{75}$$

In particular, for some constants $C, A > 0$, for $S_* \in \mathcal{S}_{r,\rho}$ and for all $t > 0$, with probability at least $1 - e^{-t}$,

$$\|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \leq C\biggl[\frac{a^2(r \wedge l)l}{n} \log\biggl(\frac{Anm}{(r \wedge l)l}\biggr) \vee \frac{\rho^2}{\lambda_{l+1}} \vee \frac{a^2 t}{n}\biggr]. \tag{76}$$

Proof. Without loss of generality, assume that $a = 1$; this would imply the general case by a simple rescaling of the problem. We will use a version of well-known bounds for least squares estimators over uniformly bounded function classes in terms of Rademacher complexities. Specifically, consider the following least squares estimator:

$$\hat{g} := \mathrm{argmin}_{g \in \mathcal{G}} \, n^{-1} \sum_{j=1}^{n} (Y_j - g(X_j))^2,$$

where $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. copies of a random couple (X, Y) in $T \times \mathbb{R}$, where $(T, \mathcal{T})$ is a measurable space, $|Y| \leq 1$ a.s. and $\mathcal{G}$ is a class of measurable functions on T uniformly bounded by 1. The goal is to estimate the regression function $g_*(x) := \mathbb{E}(Y|X = x)$. Define localized Rademacher complexity

$$\psi_n(\delta) := \mathbb{E} \sup_{G_{\mathcal{U}}, G_{\mathcal{V}} \in \mathcal{G}, \|G_{\mathcal{U}} - G_{\mathcal{V}}\|^2_{L_2(\hat{\Pi})} \leq \delta} |R_n(G_{\mathcal{U}} - G_{\mathcal{V}})|,$$

where $\hat{\Pi}$ is the distribution of X and $R_n(g) := n^{-1} \sum_{j=1}^{n} \varepsilon_j g(X_j)$ is a Rademacher process, that is $\varepsilon_1, \dots, \varepsilon_n$ is a sequence of i.i.d. Rademacher random variables independent of $\{X_j\}$. Define $\psi_n^\flat$ and $\psi_n^\sharp$ as

$$\psi_n^\flat(\delta) := \sup_{\sigma \geq \delta} \frac{\psi_n(\sigma)}{\sigma}, \qquad \psi_n^\sharp(\varepsilon) := \inf\{\delta > 0 : \psi_n^\flat(\delta) \leq \varepsilon\}.$$

The next result easily follows from Theorem 5.2 in [38]:

Proposition 3.6. There exist constants $c_1, c_2 > 0$ such that, for all $t > 0$, with probability at least $1 - e^{-t}$,

$$\|\hat{g} - g_*\|^2_{L_2(\hat{\Pi})} \le 2 \inf_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\hat{\Pi})} + c_1 \left(\psi_n^\sharp(c_2) + \frac{t}{n} \right).$$

We will apply this proposition to prove Theorem 3.5. In what follows in the proof, denote $\hat{S} := \hat{S}_l$. In our case, $T = \mathcal{V} \times \mathcal{V}$, $X = (U, V)$, and $\hat{\Pi} = \Pi^2$. Let $\mathcal{G} := \bar{\mathcal{S}}_r(l; 1)$, $g_* = S_*$ and $\hat{g} = \hat{S}$. First, we need to upper bound the Rademacher complexity $\psi_n(\delta)$ for the class $\mathcal{G}$. Let $\mathbb{S}_{r,m}(R)$ be the set of all symmetric $m \times m$ matrices S with $\mathrm{rank}(S) \le r$ and $\|S\|_F \le R$. The ε-covering number $N(\mathbb{S}_{r,m}(R); \|\cdot\|_F; \varepsilon)$ of the set $\mathbb{S}_{r,m}(R)$ with respect to the Hilbert–Schmidt distance (i.e., the minimal number of balls of radius ε needed to cover this set) can be bounded as follows:

$$N\big(\mathbb{S}_{r,m}(R); \|\cdot\|_F; \varepsilon\big) \le \left(\frac{18R}{\varepsilon} \right)^{(m+1)r}. \tag{77}$$

Such bounds are well known (see, e.g., [38], Lemma 9.3 and references therein; the proof of this lemma can be easily modified to obtain (77)). Bound (77) will be used to control the covering numbers of the set of kernels $\mathcal{S}_r(l; 1)$. Since kernels $S \in \mathcal{S}_r(l; 1)$ can be viewed as symmetric $l \times l$ matrices of rank at most $r \wedge l$ with $\|S\|_{L_2(\Pi^2)} \le 1$ and $\|S\|_F = m\|S\|_{L_2(\Pi^2)} \le m$, we conclude that the set $\mathcal{S}_r(l; 1)$ can be identified with a subset of the set $\mathbb{S}_{r\wedge l, l}(m)$. Therefore, we get the following bound:

$$N(\mathcal{S}_r(l; 1); \|\cdot\|_F; \varepsilon) \le \left(\frac{18m}{\varepsilon} \right)^{(l+1)(r\wedge l)}.$$

Since $\|S_1^1 - S_2^1\|_F^2 \le \|S_1 - S_2\|_F^2$ (truncation of the entries reduces the Hilbert–Schmidt distance), we also have

$$N(\bar{\mathcal{S}}_r(l; 1); \|\cdot\|_F; \varepsilon) \le \left(\frac{18m}{\varepsilon} \right)^{(l+1)(r\wedge l)}.$$

Let Π_n denotes the empirical distribution based on observations $(U_1, V_1), \ldots, (U_n, V_n')$. Note that,

$$\|S_1 - S_2\|^2_{L_2(\Pi_n)} = \frac{1}{n} \sum_{j=1}^n \langle S_1 - S_2, E_{U_j, V_j} \rangle^2 \le \|S_1 - S_2\|_F^2.$$

Therefore, we get the following bound on the $L_2(\Pi_n)$-covering numbers of the set $\bar{\mathcal{S}}_r(l;1)$:

$$N(\bar{\mathcal{S}}_r(l;1);L_2(\Pi_n);\varepsilon) \le \left(\frac{18m}{\varepsilon}\right)^{(l+1)(r\wedge l)}.$$

The last bound allows us to use inequality (3.17) in [38] to control the localized Rademacher complexity $\psi_n(\delta)$ of the class $\mathcal{G}$ as follows:

$$\begin{aligned}\psi_n(\delta) &= \mathbb{E}\sup_{S_1,S_2\in\bar{\mathcal{S}}_r(l;1),\|S_1-S_2\|^2_{L_2(\Pi^2)}\le\delta}\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j\big(S_1\big(U_j,V_j\big)-S_2\big(U_j,V_j\big)\big)\right| \\ &\le C_1\left[\sqrt{\frac{\delta l(r\wedge l)}{n}}\sqrt{\log\Big(\frac{Am}{\sqrt{\delta}}\Big)}\vee\frac{l(r\wedge l)}{n}\log\Big(\frac{Am}{\sqrt{\delta}}\Big)\right]\end{aligned} \tag{78}$$

with some constant $A, C_1 > 0$. This easily yields $\psi_n^\sharp(c_2) \le C_2\frac{(r\wedge l)l}{n}\log(\frac{Anm}{(r\wedge l)l})$ with some constants $A, C_2 > 0$. Proposition 3.6 now implies bound (75).

To prove bound (76), it is enough to observe that, for $S_* \in \mathcal{S}_{r,\rho}$,

$$\inf_{S\in\bar{\mathcal{S}}_r(l;1)}\|S-S_*\|^2_{L_2(\Pi^2)} \le \frac{2\rho^2}{\lambda_{l+1}}. \tag{79}$$

Indeed, since $S_* \in \mathcal{S}_{r,\rho}$, we can approximate this kernel by

$$S_l := \sum_{i,j=1}^l \langle S_*\phi_i,\phi_j\rangle(\phi_i\otimes\phi_j).$$

For the error of this approximation, we have

$$\begin{aligned}&\|S_l-S_*\|^2_{L_2(\Pi^2)} = \frac{1}{m^2}\|S_l-S_*\|_F^2 = \frac{1}{m^2}\sum_{i\vee j>l}\langle S_*\phi_i,\phi_j\rangle^2 \\ &\le \frac{1}{m}\frac{1}{\lambda_{l+1}}\sum_{i>l}\sum_{j=1}^m\lambda_i\langle S_*\phi_i,\phi_j\rangle^2 + \frac{1}{m^2}\frac{1}{\lambda_{l+1}}\sum_{i=1}^m\sum_{j>l}\lambda_j\langle S_*\phi_i,\phi_j\rangle^2 \le \frac{2\rho^2}{\lambda_{l+1}},\end{aligned}$$

which implies $\|S_l^1-S_*\|^2_{L_2(\Pi)} \le \|S_l-S_*\|^2_{L_2(\Pi^2)} \le \frac{2\rho^2}{\lambda_{l+1}}$ (since the entries of matrix S_* are bounded by 1 and truncation of the entries reduces the Hilbert–Schmidt distance). We also have $\operatorname{rank}(S_l) \le \operatorname{rank}(S_*) \le r$ and

$$\|S_l\|_{L_2(\Pi^2)} = \frac{1}{m^2}\|S_l\|_F \le \frac{1}{m^2}\|S_*\|_F = \|S_*\|_{L_2(\Pi^2)} \le \|S_*\|_{L_\infty} \le 1.$$

Therefore, $S_l^1 \in \bar{\mathcal{S}}_r(l;1)$ and bound (79) follows. Bound (76) is a consequence of (75) and (79). □

Remark 2. Note that, in the case when the basis of eigenfunctions $\{\phi_j\}$ coincides with the canonical basis of space $\mathbb{R}^{\mathcal{V}}$, the following bound holds trivially:

$$\|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \le \frac{4a^2l^2}{m^2} + \frac{2\rho^2}{\lambda_{l+1}}. \tag{80}$$

This follows from the fact that the entries of both matrices $\hat{S}_l$ and S_l are bounded by a, and their nonzero entries are only in the first l rows and the first l columns, so, $\|\hat{S}_l - S_l\|^2_{L_2(\Pi^2)} \le \frac{4a^2l^2}{m^2}$. Combining this with (76) and minimizing the resulting bound with respect to l yields the following upper bound (up to a constant) that holds for the optimal choice of l:

$$\min_{1\le l\le m}\left[\left(\frac{a^2(r\wedge l)l}{n}\log\left(\frac{Anm}{(r\wedge l)l}\right)\wedge\frac{a^2l^2}{m^2}\right)\vee\frac{\rho^2}{\lambda_{l+1}}\right]\vee\frac{a^2t}{n}.$$

It is not hard to check that, typically, this expression is of the same order (up to log factors) as the lower bound of Theorem 3.3 for $d = 1$.

3.5.2 Adaptive choice of parameters

Next we consider a penalized version of least squares estimator which is adaptive to unknown parameters of the problem (such as the rank of the target matrix and the optimal value of parameter l which minimizes the error bound of Theorem 3.5). We still assume that $|Y| \le a$ a.s. for some known constant $a > 0$. For K and A constants to be determined later, define

$$\begin{aligned}(\hat{r},\hat{l}) := \operatorname*{argmin}_{r,l=1,\dots,m}\Bigg\{&n^{-1}\sum_{j=1}^n\big(Y_j - \hat{S}_{r,l,a}\big(U_j,V_j\big)\big)^2\\ &+K\frac{a^2(r\wedge l)l}{n}\log\left(\frac{Anm}{(r\wedge l)l}\right)\Bigg\}\end{aligned} \tag{81}$$

The following theorem provides an oracle inequality for the estimator $\hat{S} := \hat{S}_{\hat{r},\hat{l},a}$..

Theorem 3.7. For a proper choice of the constants K and A in (81), there is an absolute constant C such that for all $t > 0$, probability at least $1 - e^{-t}$,

$$\begin{aligned}\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le 2\min_{1\le r\le m,1\le l\le m}\Bigg[&\inf_{S\in\bar{\mathcal{S}}_r(l;a)}\|S - S_*\|^2_{L_2(\Pi^2)}\\ &+C\left(\frac{a^2(r\wedge l)l}{n}\log\left(\frac{Anm}{(r\wedge l)l}\right)+\frac{a^2(t+\log m)}{n}\right)\Bigg].\end{aligned}$$

Proof. As in the proof of the previous theorem, we can assume that $a = 1$; the general case follows by rescaling. We will use oracle inequalities in abstract penalized empirical risk minimization problems; see [38], Theorem 6.5. We only sketch the proof here skipping the details that are standard.

As in the proof of Theorem 3.5, first consider i.i.d. copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of a random couple (X, Y) in $T \times \mathbb{R}$, where $(T, \mathcal{T})$ is a measurable space and $|Y| \le 1$ a.s. Let $\{\mathcal{G}_k : k \in I\}$ be a finite family of classes of measurable functions from T into $[-1, 1]$. Consider the corresponding family of least squares estimators

$$\hat{g}_k := \operatorname{argmin}_{g \in \mathcal{G}_k} n^{-1} \sum_{j=1}^{n} (Y_j - g(X_j))^2, k \in I.$$

Suppose the following upper bounds on localized Rademacher complexities for classes $\mathcal{G}_k, k \in I$ hold:

$$\mathbb{E} \sup_{G_{\mathcal{U}}, G_{\mathcal{V}} \in \mathcal{G}_k, \|G_{\mathcal{U}} - G_{\mathcal{V}}\|^2_{L_2(\Pi)} \le \delta} |R_n(G_{\mathcal{U}} - G_{\mathcal{V}})| \le \psi_{n,k}(\delta), \delta > 0,$$

where $\psi_{n,k}$ are nondecreasing functions of δ that do not depend on the distribution of (X, Y). Define

$$\hat{k} := \operatorname*{argmin}_{k \in I} \left[n^{-1} \sum_{j=1}^{n} \big(Y_j - \hat{g}_k(X_j)\big)^2 + K\left(\psi^{\sharp}_{n,k}(c_1) + \frac{t_k}{n}\right)\right], \tag{82}$$

Let K and c_1 be constants and let $\{t_k, k \in I\}$ be positive numbers.

We are interested in the penalized least squares estimator $\hat{g} := \hat{g}_{\hat{k}}$ of the regression function g_*. The next result is well known; it can be deduced, for instance, from Theorem 6.5 in [38].

Proposition 3.8. There exists constants $K, c_1 > 0$ in the definition (82) of $\hat{k}$ and a constant $K_1 > 0$ such that, for all $t_k > 0$, with probability at least $1 - \sum_{k \in I} e^{-t_k}$

$$\|\hat{g} - g_*\|^2_{L_2(\Pi)} \le 2 \inf_{k \in I} \left[\inf_{g \in \mathcal{G}_k} \|g - g_*\|^2_{L_2(\Pi)} + K_1 \left(\psi^{\sharp}_{n,k}(c) + \frac{t_k}{n}\right)\right].$$

We apply this result to the estimator $\hat{S} = \hat{S}_{\hat{r},\hat{l},1}$, where $(\hat{r}, \hat{l})$ is defined by (81) (with $a = 1$). In this case, $T = \mathcal{V} \times \mathcal{V}$, $X = (U, V)$, $g_* = S_*$, $I = \{(r, l) : 1 \le r, l \le$

$m\}$, $\mathcal{G}_{r,l} = \bar{\mathcal{S}}_r(l;1)$. In view of (78), we can use the following bounds on localized Rademacher complexities for these function classes:

$$\psi_{n,r,l}(\delta) := C_1\left[\sqrt{\frac{\delta l(r\wedge l)}{n}}\sqrt{\log\left(\frac{Am}{\sqrt{\delta}}\right)} \vee \frac{l(r\wedge l)}{n}\log\left(\frac{Am}{\sqrt{\delta}}\right)\right]$$

with some constant C_1, and we have

$$\psi^{\sharp}_{n,r,l}(c_1) \le C_2\frac{(r\wedge l)l}{n}\log(\frac{Anm}{(r\wedge l)l})$$

with some constant $C_2 > 0$. Define $t_{r,l} := t + 2\log m, (r,l) \in I$. This yields the bound

$$\sum_{(r,l)\in I} e^{-t_{r,l}} \le e^{-t}.$$

These considerations and Proposition 3.8 imply the claim of the theorem. □

It follows from Theorem 3.7 that, for some constant $C > 0$ and for all $t > 0$,

$$\sup_{P\in\mathcal{P}_{r,\rho,a}} \mathbb{P}_P\left\{\|\hat{S} - S_P\|^2_{L_2(\Pi^2)} \ge C\left(\Delta_n(r,\rho,a) \vee \frac{a^2 t}{n}\right)\right\} \le e^{-t}, \tag{83}$$

where

$$\Delta_n(r,\rho,a) := \min_{1\le l\le m}[\frac{a^2(r\wedge l)l}{n}\log(\frac{Anm}{(r\wedge l)l}) \vee \frac{\rho^2}{\lambda_{l+1}}].$$

Denoting

$$\tilde{l} := \min\left\{l = 1,\dots,m : (r\vee l)l\lambda_{l+1}\log\left(\frac{Anm}{(r\wedge l)l}\right) \ge \frac{\rho^2 n}{a^2}\right\},$$

it is easy to see that $\Delta_n(r,\rho,a) = \frac{a^2(r\wedge\tilde{l})\tilde{l}}{n}\log\left(\frac{Anm}{(r\wedge\tilde{l})\tilde{l}}\right) \vee \frac{\rho^2}{\lambda_{\tilde{l}}}$.

Example. Suppose that, for some $\beta > 1/2$, $\lambda_l \asymp l^{2\beta}, l = 1,\dots,m$. Under this assumption, it is easy to show that the upper bound on the squared $L_2(\Pi^2)$-error of the estimator $\hat{S}$ is of the order

$$\left(\left(\frac{a^2\rho^{1/\beta}r}{n}\log\frac{Anm}{r}\right)^{2\beta/(2\beta+1)} \wedge \left(\frac{a^2\rho^{2/\beta}\log(Anm)}{n}\right)^{\beta/\beta+1} \wedge \frac{a^2 rm\log(Anm)}{n}\right) \vee \frac{a^2 t}{n}$$

In fact, the log factors can be written in a slightly better, but more complicated way. Up to the log factors, this is the same error rate as in the lower bounds of Section 3.3; see (61).

3.6 Combining nuclear norm and squared Sobolev norm

The main goal in this section is to study the following penalized least squares estimator with a combination of two convex penalties:

$$\hat{S}_{\varepsilon,\bar{\varepsilon}} := \operatorname*{argmin}_{S\in\mathbb{D}} \left[\frac{1}{n}\sum_{j=1}^{n}\big(Y_j - S\big(U_j,V_j\big)\big)^2 + \varepsilon\|S\|_* + \bar{\varepsilon}\Big\|W^{1/2}S\Big\|^2_{L_2(\Pi^2)} \right], \quad (84)$$

where $\varepsilon, \bar{\varepsilon} > 0$ are regularization parameters, and $\mathbb{D} \subset \mathcal{S}_V$ is a closed convex set of symmetric kernels with bounded entries. That is, for all $S \in \mathbb{D}$,

$$\|S\|_{L_\infty} := \max_{u,v\in V} |S(u,v)| \le a,$$

The first penalty involved in (84) is based on the nuclear norm $\|S\|_*$, and it is used to "promote" low-rank solutions. The second penalty is based on a "Sobolev type norm" $\|W^{1/2}S\|^2_{L_2(\Pi^2)}$ and It is used to "promote" the smoothness of the solution on the graph.

We will derive an upper bound on the error $\|\hat{S}_{\varepsilon,\bar{\varepsilon}} - S_*\|^2_{L_2(\Pi^2)}$ of estimator $\hat{S}_{\varepsilon,\bar{\varepsilon}}$ in terms of spectral characteristics of the target kernel S_* and matrix W. As before, W is a nonnegatively definite symmetric kernel with spectral representation $W = \sum_{k=1}^m \lambda_k(\phi_k \otimes \phi_k)$, where $0 \le \lambda_1 \le \cdots \le \lambda_m$ are the eigenvalues of W repeated with their multiplicities and $\phi_1, \ldots, \phi_m$ are the corresponding orthonormal eigenfunctions. We will also use the decomposition of identity associated with W:

$$E(\lambda) := \sum_{\lambda_j \le \lambda} (\phi_j \otimes \phi_j), \lambda \ge 0.$$

Clearly, $\lambda \mapsto E(\lambda)$ is a nondecreasing projector-valued function. Despite the fact that the eigenfunctions $\{\phi_k\}$ are not uniquely defined in the case when W has multiple eigenvalues, the decomposition of identity $\{E(\lambda), \lambda \ge 0\}$ is uniquely defined (in fact, it can be rewritten in terms of spectral projectors of W). The distribution of the eigenvalues of W is characterized by the following *spectral function*:

$$F(\lambda) := \operatorname{tr}\big(E(\lambda)\big) = \Big\|E(\lambda)\Big\|_F^2 = \sum_{j=1}^m I(\lambda_j \le \lambda), \qquad \lambda \ge 0.$$

Denote $k_0 := F(0) + 1$ (in other words, k_0 is the smallest k such that $\lambda_k > 0$). We also assume that there exists a constant $c \geq 1$ such that $\lambda_{k+1} \leq c\lambda_k$ for all $k \geq k_0$.

In what follows, we use a regularized majorant of spectral function F. Let $\bar{F} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a nondecreasing function such that $F(\lambda) \leq \bar{F}(\lambda), \lambda \geq 0$, the function $\lambda \mapsto \frac{\bar{F}(\lambda)}{\lambda}$ is nonincreasing and, for some $\gamma \in (0,1)$,

$$\int_\lambda^\infty \frac{\bar{F}(s)}{s^2}\, ds \leq \frac{1}{\gamma}\frac{\bar{F}(\lambda)}{\lambda}, \qquad \lambda > 0.$$

Without loss of generality, we assume in what follows that $\bar{F}(\lambda) = m, \lambda \geq \lambda_m$. When that is not the case, we can take the function $\bar{F}(\lambda) \wedge m$ instead. The conditions on $\bar{F}$ are satisfied if for some $\gamma \in (0,1)$, the function $\frac{\bar{F}(\lambda)}{\lambda^{1-\gamma}}$ is nonincreasing: in this case, $\frac{\bar{F}(\lambda)}{\lambda}$ is also nonincreasing and

$$\int_\lambda^\infty \frac{\bar{F}(s)}{s^2}\, ds = \int_\lambda^\infty \frac{\bar{F}(s)}{s^{1-\gamma}}\frac{ds}{s^{1+\gamma}} \leq \frac{\bar{F}(\lambda)}{\lambda^{1-\gamma}}\int_\lambda^\infty \frac{ds}{s^{1+\gamma}} = \frac{1}{\gamma}\frac{\bar{F}(\lambda)}{\lambda}.$$

Let $S \in \mathcal{S}_V$ be a kernel that will play the role of an oracle in our analysis. Consider its spectral representation: $S = \sum_{k=1}^r \mu_k(\psi_k \otimes \psi_k)$, where $r = \operatorname{rank}(S) \geq 1$, μ_k are nonzero eigenvalues of S (possibly repeated) and ψ_k are the corresponding orthonormal eigenfunctions. Denote the range of S by L. The following *coherence function* will be used to characterize the relationship between the kernels S and W:

$$\varphi(S;\lambda) := \big\langle P_L, E(\lambda)\big\rangle := \sum_{\lambda_j \leq \lambda} \|P_L\phi_j\|^2, \qquad \lambda \geq 0. \tag{85}$$

It is immediate from this definition that $\varphi(S,\lambda) \leq F(\lambda) \leq \bar{F}(\lambda), \lambda \geq 0$. Note also that $\varphi(S;\lambda)$ is a nondecreasing function of λ and

$$\varphi(S,\lambda) = \sum_{j=1}^m \|P_L\phi_j\|^2 = r, \lambda \geq \lambda_m$$

For $\lambda < \lambda_m$, $\varphi(S;\lambda)$ can be interpreted as a "partial rank" of S. As in the case of spectral function S, we need a regularized majorant for the coherence function $\varphi(S;\lambda)$. Denote by $\Psi = \Psi_{S,W}$ the set of all nondecreasing functions $\varphi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that $\lambda \mapsto \frac{\varphi(\lambda)}{\bar{F}(\lambda)}$ is nonincreasing and $\varphi(S;\lambda) \leq \varphi(\lambda), \lambda \geq 0$. It is easy to see that

the class of functions $\Psi_{S,W}$ contains the smallest function (uniformly in $\lambda \geq 0$) that will be denoted by $\bar{\varphi}(S;\lambda)$ and it is given by the following expression:

$$\bar{\varphi}(S;\lambda) := \sup_{\sigma \leq \lambda} \bar{F}(\sigma) \sup_{\sigma' \geq \sigma} \frac{\varphi(S;\sigma')}{\bar{F}(\sigma')}.$$

It easily follows from this definition that $\bar{\varphi}(S,\lambda) = r, \lambda \geq \lambda_m$. Note that since the function $\frac{\bar{\varphi}(S,\lambda)}{\bar{F}(\lambda)}$ is nonincreasing and it is equal to $\frac{r}{m}$ for $\lambda \geq \lambda_m$, we have

$$\bar{\varphi}(S;\lambda) \geq \frac{r}{m}\bar{F}(\lambda) \geq \frac{r}{m}F(\lambda), \qquad \lambda \geq 0. \tag{86}$$

Given $t > 0$, $\tilde{\lambda} \in (0, \lambda_{k_0}]$, let $t_{n,m} := t + 3\log(2\log_2 n + \frac{1}{2}\log_2 \frac{\lambda_m}{\tilde{\lambda}} + 2)$. Suppose that, for some $D > 0$,

$$\varepsilon \geq Da\biggl(\sqrt{\frac{\log(2m)}{nm}} \vee \frac{\log(2m)}{n}\biggr). \tag{87}$$

Theorem 3.9. There exists constants C, D depending only on c, γ such that, for all $\bar{\varepsilon} \in [0, \tilde{\lambda}^{-1}]$ with probability at least $1 - e^{-t}$,

$$\begin{aligned} \|\hat{S}_{\varepsilon,\bar{\varepsilon}} - S_*\|^2_{L_2(\Pi^2)} \leq \inf_{S \in \mathbb{D}} \bigl[\|S - S_*\|^2_{L_2(\Pi^2)} \\ + Cm^2\varepsilon^2\bar{\varphi}\bigl(S;\bar{\varepsilon}^{-1}\bigr) + \bar{\varepsilon}\bigl\|W^{1/2}S\bigr\|^2_{L_2(\Pi^2)}\bigr] + C\frac{a^2 t_{n,m}}{n}. \end{aligned} \tag{88}$$

Remarks 3. Under the additional assumption that $m\log(2m) \leq n$, one can take $\varepsilon = Da\sqrt{\frac{\log(2m)}{nm}}$. In this case, the main part of the random error term in the right-hand side of bound (88) becomes

$$\begin{aligned} &Cm^2\varepsilon^2\bar{\varphi}\bigl(S;\bar{\varepsilon}^{-1}\bigr) + \bar{\varepsilon}\bigl\|W^{1/2}S\bigr\|^2_{L_2(\Pi^2)} \\ &= C'\frac{a^2\bar{\varphi}(S;\bar{\varepsilon}^{-1})m\log(2m)}{n} + \bar{\varepsilon}\bigl\|W^{1/2}S\bigr\|^2_{L_2(\Pi^2)}. \end{aligned}$$

Note also that Theorem 3.9 holds in the case when $\bar{\varepsilon} = 0$. In this case, our method coincides with nuclear norm penalized least squares (matrix LASSO) and $\bar{\varphi}(S;\bar{\varepsilon}^{-1}) = \operatorname{rank}(S)$, so the bound of Theorem 3.9 becomes

$$\begin{aligned} \|\hat{S}_{\varepsilon,0} - S_*\|^2_{L_2(\Pi^2)} \leq \inf_{S \in \mathbb{D}} \bigl[\|S - S_*\|^2_{L_2(\Pi^2)} + \\ Cm^2\varepsilon^2\operatorname{rank}(S)\bigr] + C\frac{a^2 t_{n,m}}{n}. \end{aligned} \tag{89}$$

Similar oracle inequalities were proved in [40] for a linearized least squares method with nuclear norm penalty.

Using simple aggregation techniques, it is easy to construct an adaptive estimator for which the oracle inequality of Theorem 3.9 holds with the optimal value of $\bar{\varepsilon}$ that minimizes the right-hand side of the bound. To this end, divide the sample $(U_1, V_1, Y_1), \dots, (U_n, V_n, Y_n)$ into two parts,

$$\begin{aligned} &\big(U_j, V_j, Y_j\big), \qquad j = 1, \dots, n' \quad \text{and} \\ &\big(U_{n'+j}, V_{n'+j}, Y_{n'+j}\big), \qquad j = 1, \dots, n - n', \end{aligned}$$

where $n' := [n/2] + 1$. The first part of the sample will be used to compute the estimators $\hat{S}_l := \hat{S}_{\varepsilon, \bar{\varepsilon}_l}$, $\varepsilon_l := \lambda_l^{-1}$, $l = k_0, \dots, m+1$; while the second part of the sample is used for model selection

$$\hat{l} := \operatorname*{argmin}_{l = k_0, \dots, m+1} \frac{1}{n - n'} \sum_{j=1}^{n-n'} \big(Y_{n'+j} - \hat{S}_l\big(X_{n'+j}, X'_{n'+j}\big)\big)^2.$$

Finally, let $\hat{S} := \hat{S}_{\hat{l}}$.

Theorem 3.10. Under the assumptions and notation of Theorem 3.9, with probability at least $1 - e^{-t}$,

$$\begin{aligned} \|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le \inf_{S \in \mathbb{D}} \Big[& 2\|S - S_*\|^2_{L_2(\Pi^2)} \\ & + C \inf_{\bar{\varepsilon} \in [0, \lambda_{k_0}^{-1}]} \big(m^2 \varepsilon^2 \bar{\varphi}\big(S; \bar{\varepsilon}^{-1}\big) + \bar{\varepsilon} \big\| W^{1/2} S \big\|^2_{L_2(\Pi^2)} \big) \Big] \\ & + C \frac{a^2 (\log(m+1) + t_{n,m})}{n}. \end{aligned}$$

Proof. The idea of aggregation result behind this theorem is rather well known; see [46], Chapter 8. The proof can be deduced, for instance, from Proposition 3.6 used in Section 3.5. Specifically, this proposition has to be applied in the case when $\mathcal{G}$ is a finite class of functions bounded by 1. Let $N := |\mathcal{G}|$. Then, for some numerical constant $C_1 > 0$

$$\psi_n(\delta) \le C_1 \left[\delta \sqrt{\frac{\log N}{n}} \vee \frac{\log N}{n} \right]$$

(see, e.g., [38], Theorem 3.5), and Proposition 3.6 easily implies that, for all $t > 0$, with probability at least $1 - e^{-t}$

$$\|\hat{g} - g_*\|^2_{L_2(\Pi)} \leq 2 \inf_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + C_2 \frac{\log N + t}{n}, \tag{90}$$

where $C_2 > 0$ is a constant. We will assume that $a = 1$ (in the general case, the result would follow by rescaling) and use bound (90), conditionally on the first part of the sample, in the case when $\mathcal{G} := \{\hat{g}_l : l = k_0, \dots, m+1\}$. Then, given $(U_j, V_j, Y_j), j = 1, \dots, n'$, with probability at least $1 - e^{-t}$,

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \leq 2 \min_{k_0 \leq l \leq m+1} \|\hat{S}_l - S_*\|^2_{L_2(\Pi)} + C_2 \frac{\log(m+1) + t}{n}. \tag{91}$$

By Theorem 3.9 (with t replaced by $t + \log(m+1)$) and the union bound, we get that, with probability at least $1 - e^{-t}$, for all $l = k_0, \dots, m+1$,

$$\begin{aligned} \|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \leq \inf_{S \in \mathbb{D}} \big[\|S - S_*\|^2_{L_2(\Pi^2)} \\ + C_3 m^2 \varepsilon^2 \bar{\varphi}\big(S; \bar{\varepsilon}_l^{-1}\big) + \bar{\varepsilon}_l \big\| W^{1/2} S \big\|^2_{L_2(\Pi^2)} \big] + C_3 \frac{\log(m+1) + t_{n,m}}{n} \end{aligned} \tag{92}$$

with some constant $C_3 > 0$. Therefore, we can bound

$$\min_{k_0 \leq l \leq m+1} \|\hat{S}_l - S_*\|^2_{L_2(\Pi)}$$

with the same probability by the minimum over $l = k_0, \dots, m+1$ of the expression in the right-hand side of (92). Moreover, using monotonicity of the function $\lambda \mapsto \varphi(S; \lambda)$ and the condition that $\lambda_{l+1} \leq c\lambda_l, l = k_0, \dots, m-1$, it is easy to replace the minimum over l by the infimum over $\bar{\varepsilon}$. Combining the resulting bound with (91) and adjusting the constants yields the claim. □

Using more sophisticated aggregation methods (e.g., such as the methods studied in [26]) it is possible to construct an estimator $\hat{S}$ for which the oracle inequality similar to (90) holds with constant 1 in front of the approximation error term $\|S - S_*\|^2_{L_2(\Pi^2)}$.

To understand better the meaning of function $\bar{\varphi}$ involved in the statements of Theorems 3.9 and 3.10, it makes sense to relate it to the low coherence assumptions discussed in section 3.2. Indeed, suppose that, for some $\nu = \nu(S) \geq 1$,

$$\|P_L\phi_k\|^2 \le \frac{\nu r}{m}, \qquad k = 1, \dots, m. \tag{93}$$

This is a part of standard low coherence assumptions on matrix S with respect to the orthonormal basis $\{\phi_k\}$. Clearly, it implies that

$$\bar{\varphi}(S;\lambda) \le \frac{\nu r \bar{F}(\lambda)}{m}, \qquad \lambda \ge 0. \tag{94}$$

Suppose that $n \ge m\log(2m)$ and $\varepsilon = Da\sqrt{\frac{\log(2m)}{nm}}$. If condition (94) holds for the target kernel S_* with $r = \mathrm{rank}(S_*)$ and some $\nu \ge 1$, then Theorem 3.9 implies that with probability at least $1 - e^{-t}$,

$$\|\hat{S}_{\varepsilon,\bar{\varepsilon}} - S_*\|^2_{L_2(\Pi^2)} \le C\frac{a^2\nu r\bar{F}(\bar{\varepsilon}^{-1})\log(2m)}{n} + \bar{\varepsilon}\left\|W^{1/2}S_*\right\|^2_{L_2(\Pi^2)} + C\frac{a^2 t_{n,m}}{n},$$

and Theorem 3.10 implies that with the same probability,

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le C \inf_{\bar{\varepsilon}\in[0,\lambda_{k_0}^{-1}]}\left(\frac{a^2\nu r\bar{F}(\bar{\varepsilon}^{-1})\log(2m)}{n} + \bar{\varepsilon}\left\|W^{1/2}S_*\right\|^2_{L_2(\Pi^2)}\right) + C\frac{a^2(\log(m+1)+t_{n,m})}{n}.$$

Example. If $\lambda_k \asymp k^{2\beta}$ for some $\beta > 1/2$, then it is easy to check that $\bar{F}(\lambda) \asymp \lambda^{1/2\beta}$. Under the assumption that $\|W^{1/2}S_*\|^2_{L_2(\Pi^2)} \le \rho^2$, we get the bound

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le C\left(\left(\left(\frac{a^2\rho^{1/\beta}\nu r\log(2m)}{n}\right)^{2\beta/(2\beta+1)} \wedge \frac{a^2 rm}{n}\right) \vee \frac{a^2(\log(m+1)+t_{n,m})}{n}\right). \tag{95}$$

Under the following slightly modified version of low coherence assumption (94),

$$\bar{\varphi}(S;\lambda) \le \frac{\nu(r \wedge \bar{F}(\lambda))\bar{F}(\lambda)}{m}, \qquad \lambda \ge 0, \tag{96}$$

one can almost recover upper bounds of Section 3.5,

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le C\left(\left(\left(\frac{\nu a^2\rho^{1/\beta} r\log(2m)}{n}\right)^{2\beta/(2\beta+1)} \wedge \left(\frac{\nu a^2\rho^{2/\beta}\log(2m)}{n}\right)^{\beta/(\beta+1)} \wedge \frac{a^2 rm}{n}\right) \vee \frac{a^2(\log(m+1)+t_{n,m})}{n}\right).$$

The main difference with what was proved in Section 3.5 is that now the low coherence constant ν is involved in the bounds, so the methods discussed in this section yield correct (up to log factors) error rates provided that the target kernel S_* has "low coherence" with respect to the basis of eigenfunctions of W.

Proof of Theorem 3.9. Bound (88) will be proved for a fixed oracle $S \in \mathbb{D}$ and an arbitrary function $\varphi \in \Psi_{S,W}$ with $\varphi(\lambda) = r, \lambda \geq \lambda_m$ instead of $\bar{\varphi}$. It then can be applied to the function $\bar{\varphi}$ (which is the smallest function in $\Psi_{S,W}$). Without loss of generality, we assume that $a = 1$; the general case then follows by a simple rescaling. Finally, we will denote $\hat{S} := \hat{S}_{\varepsilon,\bar{\varepsilon}}$ throughout the proof.

For a subspace L of $\mathbb{R}^{\mathcal{V}}$, let P_L be the orthogonal projection to L. We define the following orthogonal projectors $\mathcal{P}_L, \mathcal{P}_L^{\perp}$ in the space $\mathcal{S}_V$ with Hilbert–Schmidt inner product:

$$\mathcal{P}_L(A) := A - P_{L^{\perp}} A P_{L^{\perp}}, \qquad \mathcal{P}_L^{\perp}(A) = P_{L^{\perp}} A P_{L^{\perp}}, A \in \mathcal{S}_V.$$

We will use a well known representation of subdifferential of convex function $S \mapsto \|S\|_*$:

$$\partial \|S\|_* = \big\{\mathrm{sign}(S) + \mathcal{P}_L^{\perp}(M) : M \in \mathcal{S}_V, \|M\| \leq 1\big\},$$

where L is the range of S; see [38], Appendix A.4 and references therein. Denote

$$L_n(S) := \frac{1}{n}\sum_{j=1}^{n}\big(Y_j - S\big(U_j, V_j\big)\big)^2 + \varepsilon\|S\|_* + \bar{\varepsilon}\big\|W^{1/2}S\big\|^2_{L_2(\Pi^2)},$$

so that $\hat{S} := \mathrm{argmin}_{S\in\mathbb{D}} L_n(S)$. An arbitrary matrix $A \in \partial L_n(\hat{S})$ can be represented as

$$A = \frac{2}{n}\sum_{i=1}^{n}\hat{S}\big(U_i, V_i'\big)E_{U_i,V_i'} - \frac{2}{n}\sum_{i=1}^{n} Y_i E_{U_i,V_i'} + \varepsilon\hat{V} + 2\frac{\bar{\varepsilon}}{m^2}W\hat{S}, \tag{97}$$

where $\hat{V} \in \partial\|\hat{S}\|_*$. Since $\hat{S}$ is a minimizer of $L_n(S)$, there exists a matrix $A \in \partial L_n(\hat{S})$ such that $-A$ belongs to the normal cone of $\mathbb{D}$ at the point $\hat{S}$; see [2], Chapter 2,

Corollary 6. This implies that $\langle A, \hat{S} - S\rangle \le 0$ and, in view of (97),

$$
\begin{aligned}
2P_n\big(\hat{S}(\hat{S} - S)\big) - \left\langle \frac{2}{n}\sum_{i=1}^{n} Y_i E_{U_i,V_i'}, \hat{S} - S \right\rangle + \varepsilon\langle \hat{V}, \hat{S} - S\rangle \\
+ 2\frac{\bar{\varepsilon}}{m^2}\langle W\hat{S}, \hat{S} - S\rangle \le 0.
\end{aligned}
\tag{98}
$$

Here and in what follows P_n denotes the empirical distribution based on the sample $(U_1, V_1, Y_1), \dots, (U_n, V_n, Y_n)$. The corresponding true distribution of (U, V, Y) will be denoted by P. It easily follows from (98) that

$$
\begin{aligned}
2\langle \hat{S} - S_*, \hat{S} - S\rangle_{L_2(P_n)} - 2\langle \Xi, \hat{S} - S\rangle \\
+\varepsilon\langle \hat{V}, \hat{S} - S\rangle + 2\bar{\varepsilon}\big\langle W^{1/2}\hat{S}, W^{1/2}(\hat{S} - S)\big\rangle_{L_2(\Pi^2)} \le 0,
\end{aligned}
$$

where

$$
\Xi := \frac{1}{n}\sum_{j=1}^{n} \xi_j E_{U_j,V_j}, \qquad \xi_j := Y_j - S_*(U_j, V_j).
$$

We can now rewrite the last bound as

$$
\begin{aligned}
2\langle \hat{S} - S_*, \hat{S} - S\rangle_{L_2(P)} + \varepsilon\langle \hat{V}, \hat{S} - S\rangle + 2\bar{\varepsilon}\big\langle W^{1/2}(\hat{S} - S), W^{1/2}(\hat{S} - S)\big\rangle_{L_2(\Pi^2)} \\
\le -2\bar{\varepsilon}\big\langle W^{1/2}S, W^{1/2}(\hat{S} - S)\big\rangle_{L_2(\Pi^2)} + 2\langle \Xi, \hat{S} - S\rangle \\
+2(P - P_n)\big((\hat{S} - S_*)(\hat{S} - S)\big)
\end{aligned}
$$

and use a simple identity

$$
\begin{aligned}
2\langle \hat{S} - S_*, \hat{S} - S\rangle_{L_2(P)} = 2\langle \hat{S} - S_*, \hat{S} - S\rangle_{L_2(\Pi^2)} \\
= \|\hat{S} - S_*\|^2_{L_2(\Pi^2)} + \|\hat{S} - S\|^2_{L_2(\Pi^2)} - \|S - S_*\|^2_{L_2(\Pi^2)}
\end{aligned}
$$

to get the following bound:

$$
\begin{aligned}
\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} + \|\hat{S} - S\|^2_{L_2(\Pi^2)} + 2\bar{\varepsilon}\big\|W^{1/2}(\hat{S} - S)\big\|^2_{L_2(\Pi^2)} + \varepsilon\langle \hat{V}, \hat{S} - S\rangle \\
\le \|S - S_*\|^2_{L_2(\Pi^2)} - 2\bar{\varepsilon}\big\langle W^{1/2}S, W^{1/2}(\hat{S} - S)\big\rangle_{L_2(\Pi^2)} + 2\langle \Xi, \hat{S} - S\rangle \\
+2(P - P_n)(S - S_*)(\hat{S} - S) + 2(P - P_n)(\hat{S} - S)^2
\end{aligned}
\tag{99}
$$

For an arbitrary $V \in \partial\|S\|_*$, $V = \operatorname{sign}(S) + \mathcal{P}_L^{\perp}(M)$, where M is a matrix with $\|M\| \le 1$. It follows from the trace duality property that there exists an M with $\|M\| \le 1$ (to be specific, $M = \operatorname{sign}(\mathcal{P}_L^{\perp}(\hat{S}))$) such that

$$
\big\langle \mathcal{P}_L^{\perp}(M), \hat{S} - S\big\rangle = \big\langle M, \mathcal{P}_L^{\perp}(\hat{S} - S)\big\rangle = \big\langle M, \mathcal{P}_L^{\perp}(\hat{S})\big\rangle = \big\|\mathcal{P}_L^{\perp}(\hat{S})\big\|_*,
$$

where the first equality is based on the fact that $\mathcal{P}_L^\perp$ is a self-adjoint operator and the second equality is based on the fact that S has support L. Using this equation and monotonicity of subdifferentials of convex functions, we get $\langle \mathrm{sign}(S), \hat{S} - S\rangle + \|\mathcal{P}_L^\perp(\hat{S})\|_* = \langle V, \hat{S} - S\rangle \le \langle \hat{V}, \hat{S} - S\rangle$. Substituting this into the left-hand side of (99), it is easy to get

$$\begin{aligned}
&\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} + \|\hat{S} - S\|^2_{L_2(\Pi^2)} + \varepsilon\big\|\mathcal{P}_L^\perp(\hat{S})\big\|_* + 2\bar{\varepsilon}\big\|W^{1/2}(\hat{S} - S)\big\|^2_{L_2(\Pi^2)} \\
&\qquad \le \|S - S_*\|^2_{L_2(\Pi^2)} - \varepsilon\big\langle \mathrm{sign}(S), \hat{S} - S\big\rangle \\
&\qquad\quad -2\bar{\varepsilon}\big\langle W^{1/2}S, W^{1/2}(\hat{S} - S)\big\rangle_{L_2(\Pi^2)} \\
&\quad +2\langle \Xi, \hat{S} - S\rangle + 2(P - P_n)(S - S_*)(\hat{S} - S) + 2(P - P_n)(\hat{S} - S)^2
\end{aligned} \tag{100}$$

We need to bound the right-hand side of (100). We start with deriving a bound on $\langle \mathrm{sign}(S), \hat{S} - S\rangle$, expressed in terms of function φ. Note that, for all $\lambda > 0$,

$$\begin{aligned}
&\big\langle \mathrm{sign}(S), \hat{S} - S\big\rangle = \sum_{k=1}^m \big\langle \mathrm{sign}(S)\phi_k, (\hat{S} - S)\phi_k\big\rangle \\
&= \sum_{\lambda_k \le \lambda} \big\langle \mathrm{sign}(S)\phi_k, (\hat{S} - S)\phi_k\big\rangle + \sum_{\lambda_k > \lambda} \Big\langle \frac{\mathrm{sign}(S)\phi_k}{\sqrt{\lambda_k}}, \sqrt{\lambda_k}(\hat{S} - S)\phi_k\Big\rangle,
\end{aligned}$$

which easily implies

$$\begin{aligned}
&\Big|\big\langle \mathrm{sign}(S), \hat{S} - S\big\rangle\Big| \\
&\le \bigg(\sum_{\lambda_k \le \lambda} \big\|\mathrm{sign}(S)\phi_k\big\|^2\bigg)^{1/2} \bigg(\sum_{\lambda_k \le \lambda} \big\|(\hat{S} - S)\phi_k\big\|^2\bigg)^{1/2} \\
&\quad + \bigg(\sum_{\lambda_k > \lambda} \frac{\|\mathrm{sign}(S)\phi_k\|^2}{\lambda_k}\bigg)^{1/2} \bigg(\sum_{\lambda_k > \lambda} \lambda_k \big\|(\hat{S} - S)\phi_k\big\|^2\bigg)^{1/2} \\
&\le \bigg(\sum_{\lambda_k \le \lambda} \|P_L\phi_k\|^2\bigg)^{1/2} \|\hat{S} - S\|_F + \bigg(\sum_{\lambda_k > \lambda} \frac{\|P_L\phi_k\|^2}{\lambda_k}\bigg)^{1/2} \big\|W^{1/2}(\hat{S} - S)\big\|_F
\end{aligned} \tag{101}$$

We will now use the following elementary lemma.

Lemma 3.11. Let $c_\gamma := \frac{c+\gamma}{\gamma}$. For all $\lambda > 0$,

$$\sum_{\lambda_k > \lambda} \frac{\|P_L\phi_k\|^2}{\lambda_k} \le c_\gamma \frac{\varphi(\lambda)}{\lambda} \quad \text{and} \quad \sum_{\lambda_k > \lambda} \frac{1}{\lambda_k} \le c_\gamma \frac{\bar{F}(\lambda)}{\lambda}.$$

Proof. Denote $H_k := \sum_{j=1}^{l} \|P_L\phi_j\|^2, k = 1, \dots, m$. Suppose that $\lambda \in [\lambda_l, \lambda_{l+1}]$ for some $l = k_0 - 1, \dots, m-1$. We will use the properties of functions $\varphi \in \Psi_{S,W}$ and $\bar{F}$. In particular, recall that the functions $\frac{\varphi(\lambda)}{\bar{F}(\lambda)}$ and $\frac{\bar{F}(\lambda)}{\lambda}$ are nonincreasing. Using these properties and the condition that $\lambda_{k+1} \le c\lambda_k, k \ge k_0$ we get

$$\begin{aligned}
\sum_{\lambda_k > \lambda} \frac{\|P_L\phi_k\|^2}{\lambda_k} &= \sum_{k=l+1}^{m-1} H_k\Big(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k+1}}\Big) + \frac{H_m}{\lambda_m} - \frac{H_l}{\lambda_{l+1}} \\
&\le \sum_{k=l+1}^{m-1} \varphi(\lambda_k)\Big(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k+1}}\Big) + \frac{\varphi(\lambda_m)}{\lambda_m} \\
&\le c\sum_{k=l+1}^{m-1} \frac{\varphi(\lambda_{k+1})}{\lambda_{k+1}^2}(\lambda_{k+1} - \lambda_k) + \frac{\varphi(\lambda_m)}{\lambda_m} \\
&\le c\int_\lambda^\infty \frac{\varphi(s)}{s^2}\,ds + \frac{\varphi(\lambda)}{\lambda} \le c\int_\lambda^\infty \frac{\varphi(s)}{\bar{F}(s)}\frac{\bar{F}(s)}{s^2}\,ds + \frac{\varphi(\lambda)}{\lambda} \\
&\le c\frac{\varphi(\lambda)}{\bar{F}(\lambda)}\int_\lambda^\infty \frac{\bar{F}(s)}{s^2}\,ds + \frac{\varphi(\lambda)}{\lambda} \le \frac{c}{\gamma}\frac{\varphi(\lambda)}{\bar{F}(\lambda)}\frac{\bar{F}(\lambda)}{\lambda} + \frac{\varphi(\lambda)}{\lambda} \\
&= \frac{c+\gamma}{\gamma}\frac{\varphi(\lambda)}{\lambda},
\end{aligned}$$

which proves the first bound. To prove the second bound, replace in the inequalities above $\|P_L\phi_k\|^2$ by 1 and $\varphi(\lambda)$ by $\bar{F}(\lambda)$. In the case when $\lambda \ge \lambda_m$, both bounds are trivial since their left-hand sides are equal to zero. □

It follows from from (101) and the first bound of Lemma 3.11 that

$$\begin{aligned}
\big|\langle \operatorname{sign}(S), \hat{S} - S\rangle\big| &\le \sqrt{\varphi(\lambda)}\|\hat{S} - S\|_F + \sqrt{c_\gamma\frac{\varphi(\lambda)}{\lambda}}\big\|W^{1/2}(\hat{S} - S)\big\|_F \\
&= m\sqrt{\varphi(\lambda)}\|\hat{S} - S\|_{L_2(\Pi^2)} + m\sqrt{c_\gamma\frac{\varphi(\lambda)}{\lambda}}\big\|W^{1/2}(\hat{S} - S)\big\|_{L_2(\Pi^2)}.
\end{aligned} \tag{102}$$

This implies the following bound:

$$\begin{aligned}
&\varepsilon\big|\langle \operatorname{sign}(S), \hat{S} - S\rangle\big| \\
&\le \varphi(\lambda)m^2\varepsilon^2 + \frac{1}{4}\|\hat{S} - S\|_{L_2(\Pi^2)}^2 + c_\gamma\frac{\varphi(\lambda)}{\lambda}\frac{m^2\varepsilon^2}{\bar{\varepsilon}} + \frac{\bar{\varepsilon}}{4}\big\|W^{1/2}(\hat{S} - S)\big\|_{L_2(\Pi^2)}^2,
\end{aligned} \tag{103}$$

where we used twice an elementary inequality $ab \le a^2 + \frac{1}{4}b^2, a, b > 0$. We will apply this bound for $\lambda = \bar{\varepsilon}^{-1}$ to get the following inequality:

$$\begin{aligned}
&\varepsilon\big|\langle \operatorname{sign}(S), \hat{S} - S\rangle\big| \\
&\le (c_\gamma + 1)\varphi\big(\bar{\varepsilon}^{-1}\big)m^2\varepsilon^2 + \frac{1}{4}\|\hat{S} - S\|_{L_2(\Pi^2)}^2 + \frac{\bar{\varepsilon}}{4}\big\|W^{1/2}(\hat{S} - S)\big\|_{L_2(\Pi^2)}^2
\end{aligned} \tag{104}$$

To bound the next term in the right-hand side of (100), note that

$$\begin{aligned}&\bar{\varepsilon}\Big|\big\langle W^{1/2}S, W^{1/2}(\hat{S}-S)\big\rangle_{L_2(\Pi^2)}\Big|\\&\leq \bar{\varepsilon}\big\|W^{1/2}S\big\|^2_{L_2(\Pi^2)} + \frac{\bar{\varepsilon}}{4}\big\|W^{1/2}(\hat{S}-S)\big\|^2_{L_2(\Pi^2)}.\end{aligned} \tag{105}$$

The main part of the proof deals with bounding the stochastic term

$$2\langle \Xi, \hat{S}-S\rangle + 2(P-P_n)(S-S_*)(\hat{S}-S) + 2(P-P_n)(\hat{S}-S)^2$$

on the right-hand side of (100). To this end, for a fixed S and S_*, define

$$\begin{aligned}f_A(y,u,v) &:= \big(y - S_*(u,v)\big)(A-S)(u,v)\\&-(S-S_*)(u,v)(A-S)(u,v) - (A-S)^2(u,v)\\&= \big(y - S(u,v)\big)(A-S)(u,v) - (A-S)^2(u,v),\end{aligned}$$

and consider the following empirical process:

$$\alpha_n(\delta_1,\delta_2,\delta_3) := \sup\Big\{\Big|(P_n-P)(f_A)\Big| : A\in \mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\},$$

where

$$\begin{aligned}&\mathcal{T}(\delta_1,\delta_2,\delta_3)\\&:= \big\{A\in\mathbb{D} : \|A-S\|_{L_2(\Pi^2)}\leq\delta_1, \big\|\mathcal{P}_L^{\perp}A\big\|_*\leq\delta_2, \big\|W^{1/2}(A-S)\big\|_{L_2(\Pi^2)}\leq\delta_3\big\}.\end{aligned}$$

Clearly, we have

$$\begin{aligned}&2\langle \Xi, \hat{S}-S\rangle + 2(P-P_n)(S-S_*)(\hat{S}-S) + 2(P-P_n)(\hat{S}-S)^2\\&\leq 2\alpha_n\big(\|\hat{S}-S\|_{L_2(\Pi^2)}, \big\|\mathcal{P}_L^{\perp}\hat{S}\big\|_*, \big\|W^{1/2}(\hat{S}-S)\big\|_{L_2(\Pi^2)}\big),\end{aligned} \tag{106}$$

and it remains to provide an upper bound on $\alpha_n(\delta_1,\delta_2,\delta_3)$ that is uniform in some intervals of the parameters $\delta_1,\delta_2,\delta_3$. That is, to prove that the norms $\|\hat{S}-S\|_{L_2(\Pi^2)}$, $\|\mathcal{P}_L^{\perp}\hat{S}\|_*$ and $\|W^{1/2}(\hat{S}-S)\|_{L_2(\Pi^2)}$ belong to these intervals with a high probability. Note that, under the assumptions that $a=1$, $|Y|\leq a$ and all the kernels are also bounded by a, the functions f_A are uniformly bounded by a numerical constant and we have $Pf_A^2\leq c_1\|A-S\|^2_{L_2(\Pi)}$ with some numerical constant $c_1>0$. Using Talagrand's concentration inequality for empirical processes we conclude that for fixed δ_1, δ_2 and

δ_3 with probability at least $1-e^{-t}$ and with some constant $c_2>0$,

$$\alpha_n(\delta_1,\delta_2,\delta_3)\le 2\mathbb{E}\alpha_n(\delta_1,\delta_2,\delta_3)+c_2\left(\delta_1\sqrt{\frac{t}{n}}+\frac{t}{n}\right)$$

We will make this bound uniform in

$$\delta_k\in[\delta_k^-,\delta_k^+],\quad \delta_k^-<\delta_k^+,\quad k=1,2,3,$$

for some intervals to be chosen later. Define $\delta_k^j:=\delta_k^+2^{-j}$, $j=0,\dots,[\log_2(\delta_k^+/\delta_k^-)]+1, k=1,2,3$ and let $\bar t:=t+\sum_{k=1}^3\log([\log_2(\delta_k^+/\delta_k^-)]+2)$. By the union bound, with probability at least $1-e^{-t}$ and for all $j_k=0,\dots,[\log_2(\delta_k^+/\delta_k^-)]+1, k=1,2,3$, $\alpha_n(\delta_1^{j_1},\delta_2^{j_2},\delta_3^{j_3})\le 2\mathbb{E}\alpha_n(\delta_1^{j_1},\delta_2^{j_2},\delta_3^{j_3})+c_2(\delta_1^{j_1}\sqrt{\frac{\bar t}{n}}+\frac{\bar t}{n})$. By monotonicity of α_n and of the right-hand side of the bound with respect to each of the variables $\delta_1,\delta_2,\delta_3$, we conclude that with the same probability and with some numerical constant $c_3>0$, for all $\delta_k\in[\delta_k^-,\delta_k^+], k=1,2,3$,

$$\alpha_n(\delta_1,\delta_2,\delta_3)\le 2\mathbb{E}\alpha_n(2\delta_1,2\delta_2,2\delta_3)+c_3\left(\delta_1\sqrt{\frac{\bar t}{n}}+\frac{\bar t}{n}\right). \tag{107}$$

To bound the expectation $\mathbb{E}\alpha_n(2\delta_1,2\delta_2,2\delta_3)$ on the right-hand side of (107), note that, by the definition of function f_A,

$$\begin{aligned}&\mathbb{E}\alpha_n(\delta_1,\delta_2,\delta_3)\\ &\le \mathbb{E}\sup\Big\{\Big|(P_n-P)(y-S)(A-S)\Big|: A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\\ &+\mathbb{E}\sup\Big\{\Big|(P_n-P)(A-S)^2\Big|: A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\end{aligned} \tag{108}$$

A standard application of symmetrization inequality followed by contraction inequality for Rademacher sums (see, e.g., [38], Chapter 2) yields

$$\begin{aligned}&\mathbb{E}\sup\Big\{\Big|(P_n-P)(A-S)^2\Big|: A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\\ &\le 16\mathbb{E}\sup\Big\{\Big|R_n(A-S)\Big|: A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}.\end{aligned} \tag{109}$$

It easily follows from (108) and (109) that

$$\begin{aligned}\mathbb{E}\alpha_n(\delta_1,\delta_2,\delta_3)&\le \mathbb{E}\sup\Big\{\Big|\langle\Xi_1,A-S\rangle\Big|: A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\\ &+16\mathbb{E}\sup\Big\{\Big|\langle\Xi_2,A-S\rangle\Big|: A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\end{aligned} \tag{110}$$

where

$$\Xi_1 := \frac{1}{n}\sum_{j=1}^{n}(Y_j - S(U_j, V_j))E_{U_j,V_j} - \mathbb{E}(Y - S(U,V))E_{U,V}$$
$$\Xi_2 := \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j E_{U_j,V_j}, \tag{111}$$

and $\{\varepsilon_j\}$ are i.i.d. Rademacher random variables independent of the observations $(U_1, V_1, Y_1), \ldots, (U_n, V_n', Y_n)$. We will upper bound the expectations on the right-hand side of (110), which reduces to bounding $\mathbb{E}\sup\{|\langle \Xi_i, A - S\rangle| : A \in \mathcal{T}(\delta_1, \delta_2, \delta_3)\}$ for each of the random matrices Ξ_1, Ξ_2. For $i = 1, 2$ and $A \in \mathcal{T}(\delta_1, \delta_2, \delta_3)$, we have

$$\begin{aligned}\Big|\big\langle \Xi_i, A - S\big\rangle\Big| &\le \Big|\big\langle \Xi_i, \mathcal{P}_L(A - S)\big\rangle\Big| + \Big|\big\langle \Xi_i, \mathcal{P}_L^{\perp}(A)\big\rangle\Big| \\ &\le \Big|\big\langle \mathcal{P}_L\Xi_i, A - S\big\rangle\Big| + \|\Xi_i\|\Big\|\mathcal{P}_L^{\perp}(A)\Big\|_* \\ &\le \Big|\big\langle \mathcal{P}_L\Xi_i, A - S\big\rangle\Big| + \delta_2\|\Xi_i\|.\end{aligned} \tag{112}$$

To bound the spectral norm of the stochastic matrices, we use the following simple corollary of a well-known noncommutative Bernstein inequality (see, e.g., [58]) obtained by integrating exponential tails of this inequality: let Z be a random symmetric matrix with $\mathbb{E}Z = 0$, $\sigma_Z^2 := \|\mathbb{E}Z^2\|$ and $\|Z\| \le U$ for some $M > 0$ and let $Z_1, \ldots, Z_n$ be n i.i.d. copies of Z. Then

$$\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n} Z_j\right\| \le 4\left(\sigma_Z\sqrt{\frac{\log(2m)}{n}} \vee M\frac{\log(2m)}{n}\right). \tag{113}$$

To bound $\|\Xi_1\|$, we applied the bound to random variables

$$Z_j := \big(Y_j - S\big(U_j, V_j\big)\big)E_{U_j,V_j} - \mathbb{E}\big(Y - S\big(U, V\big)\big)E_{X,X'}$$

while to bound Ξ_2, we applied the bound to i.i.d. random matrices $Z_j := \varepsilon_j E_{U_j,V_j}$. In both cases, $\|Z_j\| \le 4$ and, by a simple computation, $\sigma_{Z_j}^2 := \|\mathbb{E}Z_j^2\| \le 4/m$ (see, e.g., [38], Section 9.4), bound (113) implies that, for $i = 1, 2$,

$$\mathbb{E}\|\Xi_i\| \le 16\left[\sqrt{\frac{\log(2m)}{nm}} \vee \frac{\log(2m)}{n}\right] =: \varepsilon^*. \tag{114}$$

To control the term $|\langle \mathcal{P}_L\Xi_i, A - S\rangle|$ in bound (112), we will use the following lemma.

Lemma 3.12. For all $\delta > 0$,

$$\mathbb{E} \sup_{\|M\|_F \le \delta, \|W^{1/2}M\|_F \le 1} \left|\langle \mathcal{P}_L \Xi_i, M\rangle\right| \le 4\sqrt{2}\sqrt{c_\gamma + 1}\sqrt{\frac{1}{nm}}\delta\sqrt{\varphi\big(\delta^{-2}\big)}.$$

Proof. For all symmetric $m \times m$ matrices M,

$$\langle \mathcal{P}_L \Xi_i, M\rangle = \sum_{k,j=1}^{m} \langle \mathcal{P}_L \Xi_i, \phi_k \otimes \phi_j\rangle\langle M, \phi_k \otimes \phi_j\rangle.$$

Under the following assumptions,

$$\|M\|_F^2 = \sum_{k,j=1}^{m} \left|\langle M, \phi_k \otimes \phi_j\rangle\right|^2 \le \delta^2$$
$$\left\|W^{1/2}M\right\|_F^2 = \sum_{k,j=1}^{m} \lambda_k \left|\langle M, \phi_k \otimes \phi_j\rangle\right|^2 \le 1,$$

we conclude that

$$\sum_{k,j=1}^{m} \frac{|\langle M, \phi_k \otimes \phi_j\rangle|^2}{\lambda_k^{-1} \wedge \delta^2} \le 2,$$

and it follows

$$\begin{aligned} &\left|\langle \mathcal{P}_L \Xi_i, M\rangle\right| \\ &\le \left(\sum_{k,j=1}^{m} \big(\lambda_k^{-1} \wedge \delta^2\big)\left|\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j\rangle\right|^2\right)^{1/2} \left(\sum_{k,j=1}^{m} \frac{|\langle M, \phi_k \otimes \phi_j\rangle|^2}{\lambda_k^{-1} \wedge \delta^2}\right)^{1/2} \\ &\le \sqrt{2}\left(\sum_{k,j=1}^{m} \big(\lambda_k^{-1} \wedge \delta^2\big)\left|\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j\rangle\right|^2\right)^{1/2}. \end{aligned} \tag{115}$$

Consider the following inner product:

$$\langle M_1, M_2\rangle_w := \sum_{k,j=1}^{m} \big(\lambda_k^{-1} \wedge \delta^2\big)\langle M_1, \phi_k \otimes \phi_j\rangle\langle M_2, \phi_k \otimes \phi_j\rangle,$$

and let $\|\cdot\|_w$ be the corresponding norm. We will provide an upper bound on

$$\mathbb{E}\|\mathcal{P}_L \Xi_i\|_w = \mathbb{E}(\sum_{k,j=1}^{m} (\lambda_k^{-1} \wedge \delta^2)|\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j\rangle|^2)^{1/2}.$$

Recall that

$$\Xi_i = n^{-1}\sum_{j=1}^{n} \zeta_j E_{U_j, V_j} - \mathbb{E}(\zeta E_{U,V}),$$

where $\zeta_j = Y_j - S(U_j, V_j)$ for $i = 1.$ and $\zeta_j = \varepsilon_j$ for $i = 2$. Note that in the first case $|\zeta_j| \le 2$, while in the second case $|\zeta_j| \le 1$. Therefore,

$$\mathbb{E}\|\mathcal{P}_L \Xi_i\|_w \le \mathbb{E}^{1/2}\|\mathcal{P}_L \Xi_i\|_w^2 \le \sqrt{\frac{\mathbb{E}\zeta^2 \|\mathcal{P}_L E_{U,V}\|_w^2}{n}} \le 2\sqrt{\frac{\mathbb{E}\|\mathcal{P}_L E_{U,V}\|_w^2}{n}}. \tag{116}$$

It remains to bound $\mathbb{E}\|\mathcal{P}_L E_{U,V}\|_w^2$,

$$\begin{aligned}
\mathbb{E}\Big\|\mathcal{P}_L(E_{U,V})\Big\|_w^2 &= \mathbb{E}\sum_{k,j=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big)\Big|\big\langle \mathcal{P}_L(E_{U,V}), \phi_k \otimes \phi_j\big\rangle\Big|^2 \\
&= \sum_{k,j=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big) m^{-2} \sum_{u,v \in V} \Big|\big\langle E_{u,v}, \mathcal{P}_L(\phi_k \otimes \phi_j)\big\rangle\Big|^2 \\
&\le m^{-2} \sum_{k,j=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big)\Big\|\mathcal{P}_L(\phi_k \otimes \phi_j)\Big\|_F^2 \\
&\le 2m^{-2} \sum_{k,j=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big)\big(\|P_L\phi_k\|^2 + \|P_L\phi_j\|^2\big) \\
&= 2m^{-1}\sum_{k=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big)\|P_L\phi_k\|^2 + 2m^{-2}\sum_{k=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big)\|P_L\|_F^2 \\
&= 2m^{-1}\sum_{k=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big)\|P_L\phi_k\|^2 + 2m^{-2} r \sum_{k=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big)
\end{aligned} \tag{117}$$

Note that

$$\sum_{k=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big)\|P_L\phi_k\|^2 \le \delta^2 \sum_{\lambda_k \le \delta^{-2}} \|P_L\phi_k\|^2 + \sum_{\lambda_k > \delta^{-2}} \lambda_k^{-1}\|P_L\phi_k\|^2. \tag{118}$$

Using the first bound of Lemma 3.11, we get from (118) that

$$\sum_{k=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big)\|P_L\phi_k\|^2 \le \delta^2\varphi\big(\delta^{-2}\big) + c_\gamma \delta^2 \varphi\big(\delta^{-2}\big) = (c_\gamma + 1)\delta^2\varphi\big(\delta^{-2}\big). \tag{119}$$

We also have

$$\sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \le \sum_{\lambda_k \le \delta^{-2}} \delta^2 + \sum_{\lambda_k > \delta^{-2}} \lambda_k^{-1},$$

which, by the second bound of Lemma 3.11, implies that

$$\sum_{k=1}^m \big(\lambda_k^{-1} \wedge \delta^2\big) \le \delta^2 \bar{F}\big(\delta^{-2}\big) + c_\gamma \delta^2 \bar{F}\big(\delta^{-2}\big) \le (c_\gamma + 1)\delta^2 \bar{F}\big(\delta^{-2}\big). \tag{120}$$

Using bounds (117), (119) and (120) and the fact that $\varphi(\lambda) \ge \frac{r}{m}\bar{F}(\lambda)$, we get

$$\begin{aligned}
\mathbb{E}\Big\|\mathcal{P}_L(E_{U,V})\Big\|_w^2 &\le 2m^{-1}(c_\gamma + 1)\delta^2\varphi\big(\delta^{-2}\big) + 2m^{-2} r (c_\gamma + 1)\delta^2 \bar{F}\big(\delta^{-2}\big) \\
&\le 4m^{-1}(c_\gamma + 1)\delta^2\varphi\big(\delta^{-2}\big).
\end{aligned} \tag{121}$$

The proof follows from (115), (116) and (121). □

Let $\delta := \frac{\delta_1}{\delta_3}$. Using Lemma 3.12, we get

$$\begin{aligned}
&\mathbb{E}\sup\Big\{\Big|\langle \mathcal{P}_L\Xi_i, A-S\rangle\Big| : A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\\
&\le \mathbb{E}\sup\Big\{\Big|\langle \mathcal{P}_L\Xi_i, A-S\rangle\Big| : \|A-S\|_{L_2(\Pi^2)}\le\delta_1, \Big\|W^{1/2}(A-S)\Big\|_{L_2(\Pi^2)}\le\delta_3\Big\}\\
&= \mathbb{E}\sup\Big\{\Big|\langle \mathcal{P}_L\Xi_i, A-S\rangle\Big| : \|A-S\|_{F}\le\delta_1 m, \Big\|W^{1/2}(A-S)\Big\|_{F}\le\delta_3 m\Big\}\\
&\le \delta_3 m\mathbb{E}\sup\Big\{\Big|\langle \mathcal{P}_L\Xi_i, A-S\rangle\Big| : \|A-S\|_{F}\le\delta, \Big\|W^{1/2}(A-S)\Big\|_{L_2(\Pi^2)}\le 1\Big\}\\
&\le 4\sqrt{2}\delta_3 m\sqrt{c_\gamma+1}\sqrt{\frac{1}{nm}}\delta\sqrt{\varphi\big(\delta^{-2}\big)} = 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\frac{m}{n}}\delta_1\sqrt{\varphi\big(\delta^{-2}\big)}.
\end{aligned}$$

In the case when $\delta^2\ge\bar\varepsilon$, we get

$$\mathbb{E}\sup\Big\{\Big|\langle \mathcal{P}_L\Xi_i, A-S\rangle\Big| : A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\le 4\sqrt{2}\sqrt{c_\gamma+1}\delta_1\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}}.$$

In the opposite case, when $\delta^2<\bar\varepsilon$, we use the fact that the function $\frac{\varphi(\lambda)}{\lambda}=\frac{\varphi(\lambda)}{\bar F(\lambda)}\frac{\bar F(\lambda)}{\lambda}$ is nonincreasing. This implies that $\delta^2\varphi(\delta^{-2})\le\bar\varepsilon\varphi(\bar\varepsilon^{-1})$, and we get

$$\begin{aligned}
&\mathbb{E}\sup\Big\{\Big|\langle \mathcal{P}_L\Xi_i, A-S\rangle\Big| : A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\\
&\le 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\frac{m}{n}}\delta_1\sqrt{\varphi\big(\delta^{-2}\big)} = 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\frac{m}{n}}\delta_3\sqrt{\delta^2\varphi\big(\delta^{-2}\big)}\\
&\le 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\frac{m}{n}}\delta_3\sqrt{\bar\varepsilon\varphi\big(\bar\varepsilon^{-1}\big)} = 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\bar\varepsilon}\delta_3\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}}.
\end{aligned}$$

We can conclude that

$$\begin{aligned}
&\mathbb{E}\sup\Big\{\Big|\langle \mathcal{P}_L\Xi_i, A-S\rangle\Big| : A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\\
&\le 4\sqrt{2}\sqrt{c_\gamma+1}\delta_1\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} + 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\bar\varepsilon}\delta_3\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}}.
\end{aligned}$$

This bound will be combined with (112) and (114) toget that, for $i=1,2$,

$$\begin{aligned}
&\mathbb{E}\sup\Big\{\Big|\langle \Xi_i, A-S\rangle\Big| : A\in\mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}\\
&\le \varepsilon^*\delta_2 + 4\sqrt{2}\sqrt{c_\gamma+1}\delta_1\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} + 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\bar\varepsilon}\delta_3\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}}.
\end{aligned}$$

In view of (110), this yields the bound

$$\mathbb{E}\alpha_n(\delta_1,\delta_2,\delta_3)\le C'\varepsilon^*\delta_2 + C'\delta_1\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} + C'\sqrt{\bar\varepsilon}\delta_3\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}}$$

that holds with some constant $C' > 0$ for all δ_1, δ_2, and $\delta_3 > 0$. Using (107), we conclude that for some constants C and for all $\delta_k \in [\delta_k^-, \delta_k^+], k = 1, 2, 3$,

$$\alpha_n(\delta_1, \delta_2, \delta_3) \le C\Big[\delta_1\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} + \delta_1\sqrt{\frac{\bar t}{n}} + \frac{\bar t}{n} + \varepsilon^*\delta_2 + \sqrt{\bar\varepsilon}\delta_3\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}}\Big]$$

that holds with probability at least $1 - e^{-t}$. This yields the following upper bound on the stochastic term in (100) (see also (106)):

$$\begin{aligned} &2\langle \Xi, \hat S - S\rangle + 2(P - P_n)(S - S_*)(\hat S - S) + 2(P - P_n)(\hat S - S)^2 \\ &\quad \le 2C\Big[\|\hat S - S\|_{L_2(\Pi^2)}\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} + \|\hat S - S\|_{L_2(\Pi^2)}\sqrt{\frac{\bar t}{n}} + \frac{\bar t}{n} \\ &\qquad + \varepsilon^*\|\mathcal{P}_L\hat S\|_* + \sqrt{\bar\varepsilon}\Big\|W^{1/2}(\hat S - S)\Big\|_{L_2(\Pi^2)}\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}}\Big] \end{aligned} \tag{122}$$

that holds provided that

$$\begin{aligned} \|\hat S - S\|_{L_2(\Pi^2)} &\in \big[\delta_1^-, \delta_1^+\big], \\ \Big\|\mathcal{P}_L^\perp\hat S\Big\|_* &\in \big[\delta_2^-, \delta_2^+\big], \\ \Big\|W^{1/2}(\hat S - S)\Big\|_{L_2(\Pi^2)} &\in \big[\delta_3^-, \delta_3^+\big]. \end{aligned} \tag{123}$$

We substitute bound (122) in (100) and further bound some of its terms as follows:

$$\begin{aligned} 2C\|\hat S - S\|_{L_2(\Pi^2)}\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} &\le \frac{1}{8}\|\hat S - S\|^2_{L_2(\Pi^2)} + 8C^2\frac{m\varphi(\bar\varepsilon^{-1})}{n}, \\ 2C\|\hat S - S\|_{L_2(\Pi^2)}\sqrt{\frac{\bar t}{n}} &\le \frac{1}{8}\|\hat S - S\|^2_{L_2(\Pi^2)} + 8C^2\frac{\bar t}{n}, \\ 2C\sqrt{\bar\varepsilon}\Big\|W^{1/2}(\hat S - S)\Big\|_{L_2(\Pi^2)}\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} &\le \frac{1}{4}\bar\varepsilon\Big\|W^{1/2}(\hat S - S)\Big\|^2_{L_2(\Pi^2)} + 4C^2\frac{m\varphi(\bar\varepsilon^{-1})}{n}. \end{aligned}$$

We will also use (104) to control the term $\varepsilon|\langle \mathrm{sign}(S), \hat S - S\rangle|$ in (100) and (105) to control the term $\bar\varepsilon|\langle W^{1/2}S, W^{1/2}(\hat S - S)\rangle|$. If condition (87) holds with $D \ge 32C$, then $\varepsilon \ge 2C\varepsilon^*$. By a simple algebra, it follows from (100) that

$$\begin{aligned} \|\hat S - S_*\|^2_{L_2(\Pi^2)} \;\le\; &\|S - S_*\|^2_{L_2(\Pi^2)} + C_1m^2\varepsilon^2\varphi\big(\bar\varepsilon^{-1}\big) + C_1\frac{m\varphi(\bar\varepsilon^{-1})}{n} \\ &+ \bar\varepsilon\Big\|W^{1/2}S\Big\|^2_{L_2(\Pi^2)} + \frac{\bar t}{n} \end{aligned}$$

with some constant $C_1 > 0$. Since, under condition (87) with $a = 1$, $m^2\varepsilon^2 \geq D^2\frac{m\log(2m)}{n} \geq D^2\frac{m}{n}$, we can conclude that

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \leq \|S - S_*\|^2_{L_2(\Pi^2)} + C_2 m^2\varepsilon^2\varphi\big(\bar{\varepsilon}^{-1}\big) + \bar{\varepsilon}\big\|W^{1/2}S\big\|^2_{L_2(\Pi^2)} + \frac{\bar{t}}{n} \qquad (124)$$

with some constant $C_2 > 0$.

We still have to choose the values of δ_k^-, δ_k^+ and to handle the case when conditions (123) do not hold. Due to the assumption $\|S\|_{L_\infty} \leq 1, S \in \mathbb{D}$, we note that,

$$\|\hat{S} - S\|_{L_2(\Pi)} \leq 2,$$
$$\|\mathcal{P}_L^\perp \hat{S}\|_* \leq \|\hat{S}\|_* \leq \sqrt{m}\|\hat{S}\|_F \leq m^{3/2},$$
$$\|W^{1/2}(\hat{S} - S)\|_{L_2(\Pi^2)} \leq 2\sqrt{\lambda_m}.$$

Thus, we can set $\delta_1^+ := 2, \delta_2^+ := m^{3/2}, \delta_3^+ := 2\sqrt{\lambda_m}$, which guarantees that the upper bounds of (123) are satisfied. We will also set $\delta_1^- = \delta_2^- := n^{-1/2}, \delta_3^- := \sqrt{\frac{\tilde{\lambda}}{n}}$. In the case when one of the lower bounds of (123) does not hold, we can still use inequality (122), but we have to replace each of the norms $\|\hat{S} - S\|_{L_2(\Pi)}$, $\|\mathcal{P}_L^\perp \hat{S}\|_*$, and $\|W^{1/2}(\hat{S} - S)\|_{L_2(\Pi^2)}$ which are smaller than the corresponding δ_k^- by the quantity δ_k^-. Then it is straightforward to check that inequality (124) still holds for some value of constant $C_2 > 0$. With the above choice of δ_k^-, δ_k^+, we have

$$\bar{t} \leq t + 3\log\left(2\log_2 n + \frac{1}{2}\log_2\frac{\lambda_m}{\tilde{\lambda}} + 2\right) = t_{n,m}.$$

This completes the proof. □

REFERENCES

[1] Ahlswede, R. and Winter, A., "Strong converse for identification via quantum channels," *Information Theory, IEEE Transactions on*, vol. 48, no. 3, pp. 569–579, 2002.

[2] Aubin, J.-P. and Ekeland, I., *Applied nonlinear analysis*. Courier Dover Publications, 2006.

[3] Ayazoglu, M. and Sznaier, M., "An algorithm for fast constrained nuclear norm minimization and applications to systems identification," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 3469–3475, IEEE, 2012.

[4] Balzano, L., Nowak, R., and Recht, B., "Online identification and tracking of subspaces from highly incomplete information," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pp. 704–711, IEEE, 2010.

[5] Bauer, M. and Golinelli, O., "Random incidence matrices: moments of the spectral density," *Journal of Statistical Physics*, vol. 103, no. 1-2, pp. 301–337, 2001.

[6] Beck, A. and Teboulle, M., "Gradient-based algorithms with applications to signal recovery," *Convex Optimization in Signal Processing and Communications*, 2009.

[7] Bellogín, A. and Parapar, J., "Using graph partitioning techniques for neighbour selection in user-based collaborative filtering," in *Proceedings of the sixth ACM conference on Recommender systems*, pp. 213–216, ACM, 2012.

[8] Bhatia, R., *Matrix analysis*, vol. 169. Springer, 1997.

[9] Boyd, D. M. and Ellison, N. B., "Social network sites: definition, history, and scholarship," *Engineering Management Review, IEEE*, vol. 38, no. 3, pp. 16–31, 2010.

[10] Cai, J.-F., Candès, E. J., and Shen, Z., "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[11] Candes, E. J. and Plan, Y., "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[12] Candes, E. J. and Plan, Y., "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *Information Theory, IEEE Transactions on*, vol. 57, no. 4, pp. 2342–2359, 2011.

[13] Candès, E. J. and Recht, B., "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[14] Candès, E. J. and Tao, T., "The power of convex relaxation: Near-optimal matrix completion," *Information Theory, IEEE Transactions on*, vol. 56, no. 5, pp. 2053–2080, 2010.

[15] Chatterjee, S., "Matrix estimation by universal singular value thresholding," *arXiv preprint arXiv:1212.1247*, 2012.

[16] Dai, W. and Milenkovic, O., "Set: an algorithm for consistent matrix completion," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 3646–3649, IEEE, 2010.

[17] De la Pena, V. and Giné, E., *Decoupling: from dependence to independence*. Springer, 1999.

[18] Ding, X., Jiang, T., and others, "Spectral distributions of adjacency and laplacian matrices of random graphs," *The Annals of Applied Probability*, vol. 20, no. 6, pp. 2086–2117, 2010.

[19] Donoho, D. L. and Gavish, M., "The optimal hard threshold for singular values is 4/sqrt (3)," *arXiv preprint arXiv:1305.5870*, 2013.

[20] Eckart, C. and Young, G., "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[21] Erdős, Paul, P. and Rényi, A., "On the evolution of random graphs," *Publications of tke Matkemafical Insfifufe of the Hungarian Academy of Sciences*, vol. 5, 1960.

[22] Erdős, L., Knowles, A., Yau, H.-T., and Yin, J., "Spectral statistics of erdős-rényi graphs ii: Eigenvalue spacing and the extreme eigenvalues," *Communications in Mathematical Physics*, vol. 314, no. 3, pp. 587–640, 2012.

[23] Erdős, L., Knowles, A., Yau, H.-T., Yin, J., and others, "Spectral statistics of erdős–rényi graphs i: Local semicircle law," *The Annals of Probability*, vol. 41, no. 3B, pp. 2279–2375, 2013.

[24] Fountoulakis, K., Gondzio, J., and Zhlobich, P., "Matrix-free interior point method for compressed sensing problems," *Mathematical Programming Computation*, pp. 1–31, 2012.

[25] Foygel, R. and Srebro, N., "Concentration-based guarantees for low-rank matrix reconstruction," *arXiv preprint arXiv:1102.3923*, 2011.

[26] Gaïffas, S. and Lecué, G., "Hyper-sparse optimal aggregation," *The Journal of Machine Learning Research*, vol. 12, pp. 1813–1833, 2011.

[27] Gilbert, E. N., "Random graphs," *The Annals of Mathematical Statistics*, pp. 1141–1144, 1959.

[28] Goldfarb, D., Ma, S., and Wen, Z., "Solving low-rank matrix completion problems efficiently," in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pp. 1013–1020, IEEE, 2009.

[29] Gross, D., "Recovering low-rank matrices from few coefficients in any basis," *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1548–1566, 2011.

[30] Guy, I., Ronen, I., and Wilcox, E., "Do you know?: recommending people to invite into your social network," in *Proceedings of the 14th international conference on Intelligent user interfaces*, pp. 77–86, ACM, 2009.

[31] Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G., *Recommender systems: an introduction*. Cambridge University Press, 2010.

[32] Keshavan, R. H., Montanari, A., and Oh, S., "Matrix completion from a few entries," *Information Theory, IEEE Transactions on*, vol. 56, no. 6, pp. 2980–2998, 2010.

[33] Keshavan, R. H., Montanari, A., and Oh, S., "Matrix completion from noisy entries," *The Journal of Machine Learning Research*, vol. 99, pp. 2057–2078, 2010.

[34] Király, F. J., Theran, L., Tomioka, R., and Uno, T., "The algebraic combinatorial approach for low-rank matrix completion," *arXiv preprint arXiv:1211.4116*, 2012.

[35] Klopp, O., "Rank penalized estimators for high-dimensional matrices," *Electronic Journal of Statistics*, vol. 5, pp. 1161–1183, 2011.

[36] Klopp, O., "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[37] Koh, K., Kim, S.-J., and Boyd, S., "An interior-point method for large-scale ,Ñì 1-regularized logistic regression.," *Journal of Machine learning research*, vol. 8, no. 7, 2007.

[38] Koltchinskii, V., *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, vol. 2033. Springer, 2011.

[39] Koltchinskii, V., "Sharp oracle inequalities in low rank estimation," *arXiv preprint arXiv:1210.1144*, 2012.

[40] Koltchinskii, V., Lounici, K., and Tsybakov, A. B., "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, vol. 39, no. 5, pp. 2302–2329, 2011.

[41] Koltchinskii, V. and Rangel, P., "Low rank estimation of similarities on graphs," in *High Dimensional Probability VI*, pp. 305–325, Springer, 2013.

[42] Li, L. and Toh, K.-C., "An inexact interior point method for l 1-regularized sparse covariance selection," *Mathematical Programming Computation*, vol. 2, no. 3-4, pp. 291–315, 2010.

[43] Liu, Y.-J., Sun, D., and Toh, K.-C., "An implementable proximal point algorithmic framework for nuclear norm minimization," *Mathematical programming*, vol. 133, no. 1-2, pp. 399–436, 2012.

[44] Liu, Z. and Vandenberghe, L., "Interior-point method for nuclear norm approximation with application to system identification," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, 2009.

[45] Ma, S., Goldfarb, D., and Chen, L., "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, vol. 128, no. 1-2, pp. 321–353, 2011.

[46] Massart, P., *Concentration inequalities and model selection*, vol. 1896. Springer, 2007.

[47] Negahban, S. and Wainwright, M. J., "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1665–1697, 2012.

[48] Nesterov, Y., "A method of solving a convex programming problem with convergence rate o (1/k2)," in *Soviet Mathematics Doklady*, vol. 27-2, pp. 372–376, 1983.

[49] Nesterov, Y. and Nemirovski, A., "On first-order algorithms for l 1/nuclear norm minimization," *Acta Numerica*, vol. 22, pp. 509–575, 2013.

[50] Ngo, T. T. and Saad, Y., "Scaled gradients on grassmann manifolds for matrix completion.," in *NIPS*, pp. 1421–1429, 2012.

[51] Parikh, N. and Boyd, S., "Proximal algorithms," *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[52] Paulsen, V., *Completely bounded maps and operator algebras*, vol. 78. Cambridge University Press, 2002.

[53] Phuong, N. D., Phuong, T. M., and others, "A graph-based method for combining collaborative and content-based filtering," in *PRICAI 2008: Trends in Artificial Intelligence*, pp. 859–869, Springer, 2008.

[54] Recht, B., Fazel, M., and Parrilo, P. A., "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[55] Rohde, A. and Tsybakov, A. B., "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, vol. 39, no. 2, pp. 887–930, 2011.

[56] Shapira, B., *Recommender systems handbook*. Springer, 2011.

[57] Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[58] Tropp, J. A., "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[59] Tsybakov, A. B., *Introduction to nonparametric estimation*, vol. 11. Springer, 2009.

[60] Watson, G. A., "Characterization of the subdifferential of some matrix norms," *Linear Algebra and its Applications*, vol. 170, pp. 33–45, 1992.

[61] Wen, Z., Yin, W., and Zhang, Y., "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Mathematical Programming Computation*, vol. 4, no. 4, pp. 333–361, 2012.

[62] Yurinsky, V., "Sums and gaussian vectors, volume 1617 of lecture notes in mathematics," 1995.

VITA

Pedro Andrés Rangel was born in Bogota, Colombia. He received a Bachelor of Science in Electrical Engineering from Universidad Distrital Francisco Jose de Caldas in August 2003, and a Master of Science in Electrical Engineering from Universidad de los Andes in August 2005. In 2009, he joined the doctoral program of the School of Math at the Georgia Institute of Technology.